Rheumatoid arthritis prevalence models for small populations:

Technical Document produced for Arthritis Research UK

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

October 2018

Contents

1 Executive Summary 1
2 Background 4
2.1 RA Risk Factors 4
2.1.1 Risk factor – Obesity 4
2.1.2 Risk factor – Smoking 4
2.1.3 Risk factor – Infections 5
2.1.4 Risk factor – Blood Transfusion 5
2.1.5 Risk factor – Alcohol 5
2.1.6 Risk factor – Education 5
2.1.7 Risk factor – Occupation 5
2.1.8 Risk factor – Silica exposure 5
2.2 RA incidence from literature 9
2.3 RA prevalence from the literature 10
3 Methods 12
3.1 Validation studies of self-reported RA 12
3.2 RA prevalence from English national survey data: English Longitudinal Study of Ageing 12
3.3 RA prevalence from English national survey data: Health Survey for England (2005) data 14
3.3.1 HSfE RA outcome variable 14
3.4 RA prevalence from UK primary care data: Clinical Practice Research Datalink 15
3.4.1 Joint involvement 21
3.4.2 Serology and APR tests results 22
3.4.3 Patients with HES RA diagnosis 23
3.4.4 Patients on DMARDs without other inflammatory arthritis diagnosis 23
3.4.5 CPRD risk factors 23
3.4.6 CPRD descriptive analyses 24
3.4.7 CPRD regression modelling 24
3.4.8 Interactions 26
3.4.9 Internal validation 26
3.4.10 External validation 26
3.5 Local prevalence estimates 26
3.5.1 Method 1: bootstrapping procedure to produce repeated samples 27
3.5.2 Method 2: Logistic regression and sampling-probability weights 29
3.6 Local prevalence estimates for other UK countries 30
3.7 Validation of local estimates 30
4 Results 32
4.1 RA prevalence from English national survey data: English Longitudinal Study of Ageing 32
4.1.1 Baseline characteristics of ELSA respondents 32
4.1.2 RA prevalence in each ELSA wave 33

4.1.3 RA incidence and prevalence in ELSA (refined RA case definition; excluded if has hip OA and hip pain) 35
4.1.4 RA incidence and prevalence (refined RA case definition, excluded if has hip pain) 35
4.1.5 RA incidence and prevalence (refined RA case definition, excluded if has hip pain OR hip replacement due to arthritis) 36
4.2 RA prevalence from English national survey data using Health Survey for England (2005) 37
4.2.1 RA prevalence (based on rheumatic disease medication) 38
4.2.2 RA prevalence (based on rheumatic disease medication 40
4.2.3 RA prevalence (based on rheumatic disease medication and patient-reported RA) 40
4.3 Comparing prevalence obtained using ELSA and HSfE 2005 41
4.4 ELSA risk factor statistical analyses 42
4.4.1 Internal validation of ELSA: How good is our model at predicting RA caseness? 46
4.4.2 HSfE risk factor statistical analysis 48
4.5 CPRD RA definitions, incidence & prevalence 49
4.5.1 Data extraction 49
4.5.2 Doctor diagnosed RA cases 49
4.5.3 Algorithm identified “probable RA cases” 49
4.5.4 Additional RA cases from HES outpatient dataset 50
4.5.5 Patients on DMARDs without other inflammatory arthritis diagnosis 50
4.5.6 CPRD prevalence and incidence 50
4.5.7 Baseline comparison between doctor-diagnosed cases and algorithm-defined cases 58
4.5.8 Doctor diagnosis delays 58
4.6 Regression modelling using CPRD data 61
4.6.1 Missing data 61
4.6.2 Baseline descriptive characteristics of CPRD patients 63
4.6.3 CPRD univariate logistic analysis 65
4.6.4 Multivariate logistic analysis 65
4.6.5 ROC curves 73
4.6.6 Probability and sensitivity/specificity analysis 77
4.7 Population RA prevalence using local estimation Method 2, sampling-probability weights 78
4.8 Validation of local estimates 82
4.8.1 Internal validation of local estimates 82
4.8.2 External validation of local estimates 82
4.8.3 Bland-Altman plots 85
4.9 Production of Scottish local estimates 87
4.9.1 Methods 87
4.9.2 Results 87
4.9.3 Internal validation 89
4.10 Production of Wales local estimates 95

4.10.1 Methods 95
4.10.2 Results 97

5 References 100
6 Appendix: additional information 104

6.1 ELSA outcome and risk factor definitions 104
6.1.1 ELSA RA outcome creation process 104
6.1.2 ELSA risk factor questions 107
6.1.3 Preparing/cleaning ELSA data 112
6.1.4 Risk factors in ELSA 114
6.2 Further ELSA statistical analyses 116
6.3 ROC curves using different RA definitions 118
6.4 Health Survey for England (2005) structure 121
6.5 Further HSfE statistical analysis 123
6.6 CPRD medcodes for joint involvement 127
6.7 Number of small joints involved 128

Rheumatoid arthritis prevalence model Technical Document

1 Executive Summary

Project objectives
The original objectives of this ARUK-funded project, which is part of a larger project to develop prevalence models and other related epidemiologic tools for rheumatoid arthritis (RA) and three other musculoskeletal (MSK) diseases, were as follows:

1. To develop from nationally (England) representative survey data a prevalence model for RA
2. To apply this to English general practice and MLSOA populations
3. To project these estimates to 2021-22 using population age and other risk factor projections
4. To relate these at clinical commissioning group (CCG) and below where possible to current NHS costs and NHS-funded activity
5. To make these data available in a user-friendly format on ARUK and other national websites e.g. Public Health England, NHS Information Centre
6. Data discovery to apply the RA model to Wales, N Ireland and Scotland
7. Data discovery for external validation of the RA prevalence model
8. Use a sample of CPRD RA dataset for internal RA model validation
9. Use other survey data source for RA model external validation
10. Produce estimates of time from first RA clinical manifestation and from diagnostic algorithm being met to time of diagnosis entry
11. Undertake a geospatial comparison of observed/expected prevalence of RA
12. Acquire risk factor and small population data for all models for Wales and Scotland, fracture risk for N Ireland
13. Apply all prevalence models to data for Wales and Scotland

Some of these objectives have changed and some (such as applying the data to Wales and Scotland) have been delayed and are still in train. This Technical Document covers the work undertaken for objectives 1-5 and 8-9.

Background
The Background summarises literature reviews of RA incidence, prevalence and risk or protective factors, which include gender, obesity, smoking, infections, healthcare interventions, alcohol, educational level, occupation and associated exposures. A total of 19 prevalence studies have been published, of which only one was from the UK, The Norfolk Arthritis Register (NoAR) study. Extrapolating the NoAR data to the population of the UK yields an estimate of the overall prevalence of RA in adults of 0.81% (1.16% for women and 0.44% for men, a female:male ratio of 2.7:1). On the basis of these figures, there were around 386,600 people in the UK with adult-onset RA in the year 2000.

Methods

We investigated three possible national data sources to develop the RA model: the English Longitudinal Study of Ageing (ELSA), the Health Survey for England (HSfE), and the Clinical Practice Research Datalink (CPRD). All three data sources required the development of a diagnostic algorithm. Because of the clinical, prescribing and test data available in CPRD, this algorithm was by far the most comprehensive. Prevalence in ELSA and HSfE rely to a large extent on patient self-reports. While these have been found to be reasonably reliable for some diseases e.g. stroke, the positive predictive value of self-reported RA has been studied in various populations and was found to be low, ranging between 21% and 34%, possibly due to confusion with other forms of arthritis, such as osteoarthritis.

We fitted a range of multivariate logistic regression models for in order to obtain the best performing. we internally validated the models by generating receiver operating characteristic (ROC) curves, by using the ***predict*** regression post-estimation command to generate for each CPRD patient the probability of having back pain using the derived odds ratios (ORs), and by using these probabilities to examine sensitivity and specificity. We compared aggregated local prevalence estimates with the Regional prevalence in the training dataset. The variables included in the final model are also determined by the availability of local data to match with the model variables. Hence variable selection has to be a compromise between the best model which can be produced from CPRD data and the local variable available. So far we have externally validated the local estimates against the NoAR prevalence data. Given the lack of other similar datasets in the UK we have not been able to carry out other external validations.

Derived risk factor regression coefficients are used to estimate prevalence in small population subgroups. Matching local population breakdowns for each risk factor are used, where these are available. We have three methods to produce local estimates based on the regression modelling, one using Excel VBA code (which lacks CIs and is not pre-calculated), and two using Stata software. One uses a bootstrapping method to produce repeated samples (Method 1), the other (Method 2) uses sampling-probability weights. Both methods produce CIs for the estimates, which are derived from the variance in the logistic model, not the local populations.

Results

We compared different RA prevalence stratified by age and sex that were obtained using three different RA case definitions using both ELSA and HSfE 2005 data sources, with NoAR data. Both the HSfE and ELSA definitions which rely on patient reports greatly overestimate prevalence compared to NoAR, so they cannot be used in isolation. The HSfE definition “taking RA drugs (broader definition) with patient-reported RA” appears to give similar prevalence to NoAR, and might be more reliable. While risk factor odds ratios (ORs) from ELSA models were similar to published values, internal validation of ELSA models were suboptimal with an area under the receiver operating characteristics curve of about 0.62- 0.65. Similar results were obtained with HSfE data. For these reasons, and because the breadth of the data allowed us to identify possible cases based on clinical, test and prescribing data, CPRD was clearly the best data source to use for the prevalence model.

Using these various types of CPRD data we found 86,893 patients with doctor diagnoses of RA, 910 other patients with Hospital Episode Statistics RA discharge diagnoses, 12,762 possible/probable cases with clinical and test evidence of RA, 5,589 prescribed disease-modifying anti-rheumatic drugs (DMARDs) without RA or another indication for them, giving a total of 106,154 RA cases. Thus the other possible cases increased the overall prevalence by about 18%. On a UK basis this suggests there may be about 85,000 probable RA cases without a GP diagnosis. We undertook several analyses comparing doctor-diagnosed and clinical algorithm-diagnosed patients. For example, the mean age of doctor-diagnosed patients was 60.2 and for the clinical algorithm-diagnosed cases 57.7.

We also fitted a range of logistic regression models to an RA case and control dataset, in which there were 82,736 doctor-diagnosed RA cases, 791 HES cases, 12,762 algorithm-defined cases, 5,303 DMARDs cases and 354,306 controls. ORs were similar in the different models, and on internal validation c-statistics i.e. are a under ROC curves were similar (0.74-76). To produce the local estimates we used the model which included all the cases identified above and had a c-statistic of 0.76.

We initially used Method 1, the bootstrapping method to produce repeated samples, to produce local estimates from this derivation/training dataset. However this method had been developed for whole cohort estimation. Because the dataset had a case-control (rather than whole cohort) design we found that Method 1 over-estimated prevalence at practice level. We therefore developed Method 2, which takes account of the relative sizes of the whole cohort and case-control datasets. Using as risk factors gender, age, ethnicity, deprivation, smoking alcohol and BMI this produced a whole population prevalence of 1.9% before probability weighting and 0.84% after weighting, which is only slightly below the NoAR prevalence.

Discussion

We undertook basic internal and external validation using local RA estimates from Method 2 aggregated to Regional level compared to the Regional level in the derivation dataset, and local estimates compared to QOF registers respectively. In the CPRD dataset we used for the local estimates, we identified a total of 101,870 RA registered and possible cases. After dropping cases with a death date (N=23,904), there were 77,966 cases. The average prevalence in the aggregated local estimates is lower than that in the derivation dataset. Since the estimates are based on the prevalence of risk factors in each practice, this could occur because CPRD practices differ systematically from the other practices in each Region in terms of risk factors in their populations.

As an external validation we compared aggregated local practice-level estimates with corresponding QOF register data for England Regions. The bottom row shows the percentage difference between the local estimates and QOF registers. In general the local estimates are slightly higher than the registered prevalence, as we would expect given the model we developed. The prevalence of GP-registered plus probable/possible cases in our CPRD dataset is about 20% higher than GP-registered prevalence alone, and the average prevalence in our local estimates is 15% higher than aggregated GP registers. Comparing the local estimates with NoAR, which gave a whole population prevalence of exactly 1.00% (66/6593),[1] the estimated prevalence in the East of England is 0.86%, in between the QOF registered prevalence and the NoAR prevalence. These results are reassuring, and will be explored further in the spatial analysis noted in the original objectives. Further internal validation should also be carried out, but these results are consistent with requirements.

2 Background

Rheumatoid arthritis (RA) is a chronic autoimmune disease of unclear causality, affecting around 1% of Caucasians [2 3]. The disease causes persistent joint inflammation, irreversible joint damage and premature mortality [3]. The aetiology of rheumatoid arthritis (RA) is unclear, however, both genetic and environmental factors are thought to be contributors, [2 4-7] and 40% of the causality is thought to be due to environmental factors based on analyses of twin studies.[5]. Potential environmental factors that might trigger the disease include alcohol consumption, diet (especially red meat), exposure to cats, obesity, infections, immunization, low-level of formal education, postpartum period, psychological and hormonal factors and smoking.[2 5] Smoking is a well-established risk factor in a number of studies [5 6]. Some factors suggested to be protective against developing RA including high vitamin D intake, long-term breastfeeding, regular alcohol intake and oral contraceptives.[5]

2.1 RA Risk Factors

A rapid systematic literature search was conducted, supplemented by risk factor tables supplied by ARUK. RA risk factors are shown in the following table, with associated references (**Table 1**):

Table 1: RA risk factor list

Risk factor	References
Alcohol	[3 5 8-11]
Blood transfusion and surgical procedures	[2]
Education	[5 6]
Gender	[3]
Infections	[2]
Obesity/BMI	[2 11]
Socioeconomic and Occupational class	[6 12 13]
Reproductive history[1]	[2]
Silica exposure	[7]
Smoking	[2 4 5 8 11 14-17]
Coffee consumption	[18 19]

2.1.1 Risk factor – Obesity

A primary care-based case-control study in Norfolk, England, the Norfolk Arthritis Register (NoAR) study, found an association between obesity (BMI>30) and RA (adjusted odds ratio [OR_{adj}] 3.74, 95% CI 1.14-12.27) In the overweight (body mass index [BMI] 25.0-29.9) category no increased risk was observed.[2]

2.1.2 Risk factor – Smoking

A 2001 study by Hutchison et al compared 239 outpatients with RA with 239 controls matched for age, sex and social class.[4] A dose response relationship was observed between pack years smoked and RA. A modest relationship was determined between RA and ever having smoked (matched OR 1.81, 95% CI 1.22 to 2.19; p=0.002), but heavy smoking and RA were strongly associated (matched OR 13.54, 95% CI 2.89 to 63.38; p<0.001). Pack years were used to quantify cigarette consumption (20 cigarettes smoked daily for a duration of a year).[4] In the NoAR study Symmons et al found an increased risk of RA for people that ever smoked with an OR of 1.66, 95% CI 0.95-3.06 (even after correcting for social

[1] It is only related to females

class)[2]. Smoking was confirmed to be a risk factor in a Swedish study (OR 2.26, 95% CI 1.42-3.60) compared to never and ex-smokers [5].

Recently di Giuseppe et al analysed the Swedish Mammography prospective cohort data for women aged 54 to 89 years in terms of association between RA and smoking duration and intensity.[14] Smoking intensity was significantly associated with higher risk of developing RA. The relative risk (RR) comparing 1 to 7 cigarettes/day vs never smoking was 2.31, 95% CI 1.59-3.36) as well as smoking duration (comparing 1 to 25 years vs never smoking RR 1.60, 95% CI 1.07-2.38). RA risk decreases over time after smoking cessation e.g. respondents that quit smoking 15 years ago had a 30% lower risk of having RA compared to respondents that stopped smoking a year before (RR 0.70, 95% CI 0.24-2.02).

2.1.3 Risk factor – Infections

Symmons et al in NoAR found no evidence supporting the association between RA and self-reported history of prior infection with measles, rubella, glandular fever, or tuberculosis in a community based study [2].

2.1.4 Risk factor – Blood Transfusion

On the other hand RA cases had a history of elevated number of prior blood transfusions (OR_{adj} 3.58, 95% CI 1.46-8.81) [2]. There could be several explanations for this observation- blood transfusion might be a marker of another factor such as a surgery, blood loss, transmission of a pathogen or immunologic trigger.

2.1.5 Risk factor – Alcohol

Maxwell et al, found an inverse association with both RA risk and severity.[3] Non-drinkers had an OR of 4.17 (95% CI 3.01-5.77) compared to subjects consuming alcohol on > 10 days per month. Moreover, measures of RA severity (CRP, 28-joint DAS, pain visual analogue scale, modified HAQ and modified Larsen score) had an inverse relationship with increasing alcohol consumption. A Swedish study found a lower risk of developing RA with reported moderate alcohol consumption versus low intake consumption (OR 0.48, 95% CI 0.22-1.05) after adjusting for smoking and formal level of education in multivariate analyses.[5] Moreover, individuals with infrequent alcohol consumption had higher risk of RA (OR 4.02, 95% CI 2.14-7.54 vs. recent intake).

2.1.6 Risk factor – Education

Individuals with elementary school education (<=8 years) had a higher risk of developing RA compared to individuals with a university degree (OR 2.42, 95% CI 1.18-4.93) in multivariate analyses.[5] Lower formal education, an indicator of socioeconomic status, had an increased risk of RA compared with individuals with a university degree (RR 1.4, 95% CI 1.2-1.8) [6].

2.1.7 Risk factor – Occupation

Bengtsson et al in the Swedish EIRA study found a lower RA risk for individuals with higher non-manual occupations in contrast to employees of other types (RR 1.2, 95% CI 0.9-1.6).[6]

2.1.8 Risk factor – Silica exposure

The Swedish EIRA study also observed an increased risk for RA among silica exposed persons from exposure to stone dust or being in certain occupations or workplaces, e.g. rock drilling, or stone crushing.[7] About two thirds of silica exposure were related to the building industry – electricians, construction workers, sanitary engineers, drivers, stone masons, rock drillers, painters, brick and floor layers, gardeners.[7] Men exposed to silica had a twofold higher risk of developing RA compared to men without silica exposure. Even after adjusting for smoking patterns, silica exposure remained a risk factor. **Table 2** displays pooled and/or adjusted ORs for the risk factors of RA, compiled from the literature search.

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

Risk factor	Type of Odds Ratio	Odds Ratio	95% CI	Effect on Outcome
Gender				
Women	Adjusted[3]	1.32	[1.13-1.67]	Risk Factor
Smoking				
1-10 pack years	Matched[4]	0.80	[0.44-1.50]	NS
11-20 pack years	Matched[4]	0.55	[0.26-1.16]	NS
21-30 pack years	Matched[4]	1.76	[0.95-3.29]	NS
31-40 pack years	Matched[4]	5.72	[2.28-14.36]	Risk Factor
41-50 pack years	Matched[4]	13.54	[2.89-63.38]	Risk Factor
>50 pack years	Matched[4]	8.41	[2.45-28.84]	Risk Factor
Ever smoked	Matched[4]	1.81	[1.22-2.19]	Risk Factor
Ever smoked	Adjusted[3]	2.05	[1.70-2.47]	Risk Factor
Former and never	Adjusted for smoking, education, alcohol [5]	1.00		Reference
Regular and occasional	Adjusted for smoking, education, alcohol [5]	2.26	[1.42-3.60]	Risk Factor
Never	RR Adjusted for age, menopause status, parity, alcohol use, education, BMI [14]	1.00		Reference
Former	RR Adjusted for age, menopause status, parity, alcohol use, education, BMI [14]	1.68	[1.19-2.38]	Risk Factor
Current	RR Adjusted for age, menopause status, parity, alcohol use, education, BMI [14]	2.20	[1.58-3.04]	Risk Factor
Smoking intensity (cigarettes/day)				
Never	RR Adjusted for age, menopause status, parity, alcohol use, education, BMI [14]	1.00		Reference
1 to 7	RR Adjusted for age, menopause status, parity, alcohol use, education, BMI [14]	2.31	[1.59-3.36]	Risk Factor
8 to 14	RR Adjusted for age, menopause status, parity, alcohol use, education, BMI [14]	2.19	[1.50-3.21]	Risk Factor
>40 (median 45)	RR Adjusted for age, menopause status, parity, alcohol use, education, BMI [14]	1.46	[0.96-2.23]	NS
Smoking duration (years)				
Never	RR Adjusted for age, menopause status, parity, alcohol use, education, BMI [14]	1.00		Reference
1 to 25	RR Adjusted for age, menopause status, parity, alcohol use, education, BMI [14]	1.60	[1.07-2.38]	Risk Factor
25 to 40	RR Adjusted for age, menopause status, parity, alcohol use, education, BMI [14]	1.99	[1.40-2.82]	Risk Factor
>40 (median 45)	RR Adjusted for age, menopause status, parity, alcohol use, education, BMI [14]	2.33	[1.52-3.57]	Risk Factor
Smoking (pack-years)				

Risk factor	Type of Odds Ratio	Odds Ratio	95% CI	Effect on Outcome
Never	RR Adjusted for age, menopause status, parity, alcohol use, education, BMI [14]	1.00		Reference
1 to 5	RR Adjusted for age, menopause status, parity, alcohol use, education, BMI [14]	1.72	[1.11-2.67]	Risk Factor
6 to 13	RR Adjusted for age, menopause status, parity, alcohol use, education, BMI [14]	2.19	[1.48-3.25]	Risk Factor
14 to 22	RR Adjusted for age, menopause status, parity, alcohol use, education, BMI [14]	2.04	[1.35-3.09]	Risk Factor
>22 (median 28)	RR Adjusted for age, menopause status, parity, alcohol use, education, BMI [14]	1.82	[1.19-2.79]	Risk Factor
Obesity				
Underweight + Normal weight [<20< x < 24.9]	Adjusted for smoking and social class[2]	1.00		Reference
Overweight [25-29.9]	Adjusted for smoking and social class[2]	1.08	[0.54-2.16]	NS
Obese +Severely obese [30<x<40]	Adjusted for smoking and social class[2]	3.74	[1.14-12.27]	Risk Factor
Infections				
Measles	Adjusted for smoking and social class[2]	0.76	[0.32-1.84]	NS
Rubella	Adjusted for smoking and social class[2]	0.84	[0.40-1.77]	NS
Glandular fever	Adjusted for smoking and social class[2]	0.89	[0.33-2.44]	NS
Hepatitis	Adjusted for smoking and social class[2]	0.89	[0.14-5.63]	NS
Tuberculosis	Adjusted for smoking and social class[2]	0.51	[0.04-5.80]	NS
Healthcare interventions				
Blood transfusion	Adjusted for smoking and social class[2]	4.83	[1.29-18.07]	Risk Factors
Appendectomy	Adjusted for smoking and social class[2]	2.48	[0.82-7.46]	NS
Tonsillectomy	Adjusted for smoking and social class[2]	0.95	[0.44-2.07]	NS
Reproductive history [2]				
Miscarriage	Adjusted for social class and marital status[2]	2.17	[0.86-5.49]	NS
Parity	Adjusted for social class and marital status[2]	1.24	[0.44-3.48]	NS
Hysterectomy	Adjusted for social class and marital status[2]	2.40	[0.93-6.22]	NS
Menopausal	Adjusted for social class and marital status[2]	1.01	[0.35-2.90]	NS
Alcohol frequency (days per month)				
0	Adjusted[3]	1.00		Reference
1-5	Adjusted[3]	0.30	[0.23-0.40]	NS
6-10	Adjusted[3]	0.17	[0.12-0.23]	NS
>10	Adjusted[3]	0.15	[0.11-0.21]	NS
Regular drinker				
Yes	Crude[3]	1.00		Reference

[2] For female respondents only 1.24

Risk factor	Type of Odds Ratio	Odds Ratio	95% CI	Effect on Outcome
No	Crude[3]	3.97	[3.04-5.18]	Protective for drinkers
Regular drinker				
Yes	Adjusted for age, gender, smoking[3]	1.00		Reference
No	Adjusted for age, gender, smoking[3]	2.31	[1.73-3.07]	Protective for drinkers
Alcohol consumption pattern				
Recent	Adjusted for smoking, education, alcohol [5]	1.00		Reference
Abstainers	Adjusted for smoking, education, alcohol [5]	1.07	[0.53-2.16]	NS
Infrequent	Adjusted for smoking, education, alcohol [5]	4.02	[2.14-7.54]	Risk factor
Low	Adjusted for smoking, education, alcohol [5]	1.00		Reference
Moderate	Adjusted for smoking, education, alcohol [5]	0.48	[0.22-1.05]	NS
High	Adjusted for smoking, education, alcohol [5]	0.76	[0.32-1.84]	NS
Formal education				
<= 8 years	Adjusted for smoking, education, alcohol [5]	2.19	[1.04-4.61]	Risk factor
9-10 years	Adjusted for smoking, education, alcohol [5]	1.78	[0.83-3.78]	NS
11-12 years	Adjusted for smoking, education, alcohol [5]	1.64	[0.61-4.43]	NS
>12 years	Adjusted for smoking, education, alcohol [5]	1.23	[0.46-3.32]	NS
University degree	Adjusted for smoking, education, alcohol [5]	1.00		Reference
University degree Yes	RR adjusted for age, residential area, sex, smoking [6]	1.00		Reference
University degree No	RR adjusted for age, residential area, sex, smoking [6]	RR 1.7	[1.2-2.2]	Risk factor
Compulsory school	RR adjusted for age, residential area, sex, [6]	1.5	[1.1-2.0]	Risk factor
Vocational upper secondary school	RR adjusted for age, residential area, sex, [6]	1.6	[1.1-2.3]	Risk factor
Theoretical upper secondary school	RR adjusted for age, residential area, sex, [6]	1.3	[0.9-1.9]	NS
Other education	RR adjusted for age, residential area, sex, [6]	1.3	[1.0-1.7]	NS
University degree	RR adjusted for age, residential area, sex, [6]	1.0		Reference
Occupational class				
Not higher non-manual employees	RR adjusted for age, residential area, sex, smoking [6]	1.2	[0.9-1.6]	NS

Risk factor	Type of Odds Ratio	Odds Ratio	95% CI	Effect on Outcome
Higher non-manual employees	RR adjusted for age, residential area, sex, smoking [6]	1.0		Reference
Occupational class (groups)				
Unskilled manual workers	RR adjusted for age, residential area, sex, [6]	1.2	[0.8-1.6]	NS
Skilled manual workers	RR adjusted for age, residential area, sex, [6]	1.4	[1.0-2.1]	NS
Assistant non-manual employees	RR adjusted for age, residential area, sex, [6]	1.3	[0.9-1.8]	NS
Intermediate non-manual employees	RR adjusted for age, residential area, sex, [6]	1.1	[0.8-1.5]	NS
Higher non-manual employees	RR adjusted for age, residential area, sex, [6]	1.0		Reference
Occupational exposures				
Subjects exposed to silica overall v unexposed				
18 to 49 (age)	OR adjusted for age, residential area, smoking [7]	1.6	[0.6-4.4]	NS
50 to 70 (age)	OR adjusted for age, residential area, smoking [7]	2.7	[1.2-5.8]	RF
18 to 70 (age)	OR adjusted for age, residential area, smoking [7]	2.2	[1.2-3.9]	RF
Subjects who had worked with rock drilling or stone crushing v unexposed				
18 to 49 (age)	OR adjusted for age, residential area, smoking [7]	2.6	[0.4-18.1]	NS
50 to 70 (age)	OR adjusted for age, residential area, smoking [7]	3.3	[1.1-10.1]	RF
18 to 70 (age)	OR adjusted for age, residential area, smoking [7]	3.0	[1.2-7.6]	RF
Absence of silica exposure between RA patients				
Never smoked	OR adjusted for age, residential area [7]	1.0	-	NS
Ever smoked	OR adjusted for age, residential area [7]	1.4	[0.9-2.3]	NS
Silica exposure between RA patients				
Never smoked	Adjusted for age, residential area [7]	1.1	[0.3-4.4]	NS
Ever smoked	Adjusted for age, residential area [7]	3.7	[1.7-8.1]	RF

2.2 RA incidence from literature

A systematic review by Alamos et al identified 11 studies with estimated incidence rates of RA (

Table 3).[20]

Table 3 Incidence rates of RA worldwide in studies identified by a systematic review[3]

Publication	Country	Type of study	Total*	Male*	Female*	Population age (years)
Chan 1993	USA	Retrospective	0.3	0.2	0.5	≥18
Guillemin 1994	France	Retrospective	0.1	0.1	0.1	20-70
Symmons 1994	England	Prospective	0.2	0.1	0.3	≥16
Drosos 1997	Greece	Retrospective	0.2	0.1	0.4	≥16
Uhlig 1998	Norway	Retrospective	0.3	0.1	0.4	20-79
Kaipiainen-Seppanen 2000	Finland	Retrospective	0.3*	0.2	0.4	≥16
Riise 2000	Norway	Retrospective	0.3*	0.2	0.4	≥20
Kaipiainen-Seppanen 2001	Finland	Retrospective	0.3	0.2	0.4	≥16
Doran 2002	USA	Retrospective	0.4*	0.3	0.5	≥16
Savolainen 2003	Finland	Prospective	0.4*	0.3	0.5	≥16
Soderlin 2002	Sweden	Prospective	0.2	0.2	0.3	≥16

* Incidence(cases/10^3 inhabitants)

** Crude rates

2.3 RA prevalence from the literature

The same review identified 19 studies with estimated prevalence rates of RA (see Table 4).[20] Prevalence rates varied from 1.8% in Yugoslavia to 10.7% in the USA. Note that this paper uses a population denominator of 1,000.

Table 4: prevalence estimates of RA per 1,000 population worldwide from a systematic review[3]

Publication	Country	Type of study	Total*	Male*	Female*	Population age (years)
Pountain 1991	Oman	Cross-sectional	3.6**	16		
Hakala 1993	Finland	Retrospective	8.0**	6.1	10.0	≥16
Lau 1993	China	Cross-sectional	3.5**			≥16
Drosos 1997	Greece	Retrospective	3.5	1.9	4.5	≥16
Kvien 1997	Norway	Cross-sectional	4.4*	1.9	6.7	20-79
Cimmino 1998	Italy	Cross-sectional	3.3**	1.3	5.1	≥16
Stojacovic 1998	Yugoslavia	Cross-sectional	1.8**	0.9	2.9	≥20
Gabriel 1999	USA	Retrospective	10.7	7.4	13.7	≥35
Power 1999	Ireland	Cross-sectional	5**			
Saraux 1999	France	Cross-sectional	5.0	2.4	7.6	≥18
Simmonson 1999	Sweden	Cross-sectional	5.1**		20-74	
Riise 2000	Norway	Retrospective	4.3**	2.7	5.8	≥20
Carmona 2002	Spain	Cross-sectional	5**	2	8	≥20
Spindler 2002	Argentina	Retrospective	2.0**	0.6	3.2	≥16
Symmons 2002	England	Cross-sectional	8.5**	4.4	11.2	≥16

[3] Tables adapted from reference 20. Alamanos Y, Voulgari PV, Drosos AA. Incidence and Prevalence of Rheumatoid Arthritis, Based on the 1987 American College of Rheumatology Criteria: A Systematic Review. Seminars in Arthritis and Rheumatism 2006;36:182-88. doi: 10.1016/j.semarthrit.2006.08.006 Link: http://www.sciencedirect.com/science/article/pii/S0049017206001107.

Publication	Country	Type of study	Total*	Male*	Female*	Population age (years)
Dai 2003	China	Cross-sectional	2.8	1.4	4.1	≥16
Andrianakos 2003	Greece	Cross-sectional	7**		19	
Akar 2004	Turkey	Cross-sectional	3.6**	1.5	7.7	≥20
Guillemin 2005	France	Cross-sectional	3.1	0.9	5.1	≥18

* Prevalence (cases/10^3 inhabitants)

** Crude rates

The prevalence of RA from the NoAR study is shown in Table 5 below.[1 21] Extrapolating these data to the population of the UK yields an estimate of the overall prevalence of RA in adults of 0.81% (1.16% for women and 0.44% for men, a female:male ratio of 2.7:1). On the basis of these figures, there were around 386,600 people in the UK with adult-onset RA in the year 2000.

Table 5: prevalence of RA by age and sex from the Norfolk Arthritis Register (NoAR)

	Female age groups (yr)				Male age groups (yr)			
	16–44	**45–64**	**65–74**	**75+**	**16–44**	**45–64**	**65–74**	**75+**
Stratum sample	2799	869	439	414	–	1279	724	526
Dead/not at address	283	31	9	13	–	78	21	22
True sample size	2516	838	430	401	–	1201	703	504
Response rate (%)	79	89.5	89.8	79.1	–	80.4	86.2	81.5
Number of positive responders	173	183	145	134	–	170	127	93
Proportion assessed	0.71	0.87	0.86	0.7	–	0.76	0.87	0.84
Number with RA	3	14	11	12	–	7	8	11
Minimum RA prevalence[4] (%)	0.12	1.67	2.56	2.99	0.02	0.58	1.14	2.18
(95% CI)	(0.03, 0.35)	(0.91, 2.80)	(1.28, 4.58)	(1.55, 5.23)		(0.23, 1.20)	(0.49, 2.24)	(1.09, 3.90)
RA prevalence[5] (%)	0.15	1.87	2.84	3.77		0.72	1.32	2.7
(95% CI)	(0.03, 0.44)	(1.02, 3.13)	(1.42, 5.07)	(1.95, 6.59)		(0.29, 1.49)	(0.57, 2.60)	(1.35, 4.84)
RA prevalence[6] (%)	0.21	2.15	3.31	5.36		0.94	1.51	3.08
(95% CI)	(0.07, 0.66)	(1.28, 3.61)	(1.85, 5.92)	(3.10, 9.27)		(0.45, 1.98)	(0.76, 3.00)	(1.73, 5.51)

[4] Prevalence of RA assuming that none of the non-responders to the screening questionnaire and examination had RA. This gives a minimum prevalence but seems the most robust assumption.

[5] Prevalence of RA assuming that non-responders have same rate of RA as responders.

[6] Prevalence of RA assuming that those who declined examination had the same rate of RA as those who agreed to be examined.

3 Methods

3.1 Validation studies of self-reported RA

In theory it is possible to derive RA prevalence at least from UK national survey data. We examined data from both the English Longitudinal Study of Ageing (ELSA) and the Health Survey for England (HSfE) to determine whether this was possible. Both ELSA and HSfE rely on patient self-report of an RA diagnosis, but the validity of self-reported RA has been questioned. Data obtained from clinical examination and diagnosis, as well as radiological evidence, suggest much lower prevalence than obtained using self-reports.[22]

The positive predictive value of self-reported RA has been studied in various populations and was found to be low, ranging between 21% and 34%, possibly due to confusion with other forms of arthritis, such as osteoarthritis.[22-25]. Kvien et al showed that of 5,886 respondents (3,670 with musculoskeletal pain or stiffness) 158 patients (2.7%) reported having diagnosed RA by a doctor (n=107) and/or according to their own opinion (n=142). RA was confirmed by clinical examination in only 35 of these 158 individuals (22%, 95% CI 16-29) [25]. Star et al contacted women with self-reported RA to obtain the consent to reconfirm RA diagnosis with the doctor. The self-reported diagnosis of RA was confirmed in only 26 (21%) individuals.[24] Therefore information on arthritic conditions from self-reported epidemiological studies has to be used cautiously. Formica et al concluded that self-report for RA along with prescription of disease-modifying anti-rheumatic drugs (DMARDs) is a valid case definition for identifying clinical RA, and is sufficient for use after excluding those who report other rheumatic conditions and prednisone as their only DMARD [23]. We used this information in analysing ELSA and HSfE data.

3.2 RA prevalence from English national survey data: English Longitudinal Study of Ageing

This section presents questions related to RA in the ELSA dataset in every Wave (see also **Table 6)**. Additional detailed information about the derivation of the outcome and risk factor variables is shown in Section 6.1 ELSA outcome and risk factor definitions. For Wave 0 1998 (which were HSfE questions) we have shown all the disease categories included to demonstrate the breadth of responses available. For other Wave 0 questions (1999 and 2001) we have shown only those relevant to MSK disease. Of course all variables include negative values i.e. Value = -9 Label = Refusal, Value = -8 Label = Don't know, Value = -2 Label = Schedule not applicable, Value = -1 Label = Not applicable.

Wave 0 1998
Pos. = 698-704 **Variable =** illsm1-6 **Variable label =** Type of illness - 1st
Value = 1 Label = Cancer (neoplasm) including lumps, masses, tumours and growth
Value = 2 Label = Diabetes. Incl. Hyperglycemia
Value = 3 Label = Other endocrine/metabolic
Value = 4 Label = Mental illness/anxiety/depression/nerves (nes)
Value = 5 Label = Mental handicap
Value = 6 Label = Epilepsy/fits/convulsions
Value = 7 Label = Migraine/headaches
Value = 8 Label = Other problems of nervous system
Value = 9 Label = Cataract/poor eye sight/blindness
Value = 10 Label = Other eye complaints
Value = 11 Label = Poor hearing/deafness
Value = 12 Label = Tinnitus/noises in the ear
Value = 13 Label = Meniere's disease/ear complaints causing balance problems
Value = 14 Label = Other ear complaints
Value = 15 Label = Stroke/cerebral haemorrhage/cerebral thrombosis
Value = 16 Label = Heart attack/angina
Value = 17 Label = Hypertension/high blood pressure/blood pressure (nes)

Value = 18 Label = Other heart problems
Value = 19 Label = Piles/haemorrhoids incl. Varicose Veins in anus.
Value = 20 Label = Varicose veins/phlebitis in lower extremities
Value = 21 Label = Other blood vessels/embolic
Value = 22 Label = Bronchitis/emphysema
Value = 23 Label = Asthma
Value = 24 Label = Hayfever
Value = 25 Label = Other respiratory complaints
Value = 26 Label = Stomach ulcer/ulcer (nes)/abdominal hernia/rupture
Value = 27 Label = Other digestive complaints (stomach, liver, pancreas, bile d
Value = 28 Label = Complaints of bowel/colon (large intestine, caecum, bowel, c
Value = 29 Label = Complaints of teeth/mouth/tongue
Value = 30 Label = Kidney complaints
Value = 31 Label = Urinary tract infection
Value = 32 Label = Other bladder problems/incontinence
Value = 33 Label = Reproductive system disorders
Value = 34 Label = Arthritis/rheumatism/fibrositis
Value = 35 Label = Back problems/slipped disc/spine/neck
Value = 36 Label = Other problems of bones/joints/muscles
Value = 37 Label = Infectious and parasitic disease
Value = 38 Label = Disorders of blood and blood forming organs
Value = 39 Label = Skin complaints
Value = 40 Label = Other complaints
Value = 41 Label = Unclassifiable (no other codable complaint)
Value = 42 Label = Complaint no longer present NB Only use this code if it is a

Wave 0 1999
Pos. = 840-6 **Variable =** illsm1-6 **Variable label =** Type of illness - 1st
Value = 34 Label = Arthritis/rheumatism/fibrositis
Value = 35 Label = Back problems/slipped disc/spine/neck
Value = 36 Label = Other problems of bones/joints/muscles
Value = 42 Label = Complaint no longer present NB Only use this code if it is a

Wave 0 2001
Pos. = 1277-9 **Variable =** discode1-3 **Variable label =** Disability code (1)
Value = 51 Label = Rheumatoid arthritis
Value = 52 Label = Osteoarthritis and allied disorders
Value = 53 Label = Arthritis and rheumatism not codable above
Value = 54 Label = Knee problems
Value = 55 Label = Back and neck problems
Value = 56 Label = Other joint problems
Value = 57 Label = Acquired deformities
Value = 58 Label = Other musculo-skeletal
Value = 64 Label = Insufficient data to classify
Value = 90 Label = Irrelevant response

Wave 1
Pos. = 262-4 **Variable =** heart1-3 **Variable label =** Which types of arthritis do you have? 1st
Value = 1 Label = ... osteoarthritis?
Value = 2 Label = ... rheumatoid arthritis?
Value = 3 Label = ... some other kind of arthritis?

Table 6: variables in ELSA related to RA

Wave 0 1998	Wave 0 1999	Wave 0 2001	Wave 1 (2002/03)	Wave 2 (2004/05)	Wave 3 (2006/07)	Wave 4 (2008/09)	Wave 5 (2010/11)
illsm1, illsm2, illsm3,	illsm1, illsm2, illsm3, illsm4, illsm5, ilsm6	discode1, discode2, discode3	heart1, heart2, heart3	HeArt1, HeArt2	heartra	heartra	heartra

Wave 0 1998	Wave 0 1999	Wave 0 2001	Wave 1 (2002/03)	Wave 2 (2004/05)	Wave 3 (2006/07)	Wave 4 (2008/09)	Wave 5 (2010/11)
illsm4,[7] ***illsm5***[8]***, illsm6***[9]							
				bheart1, bheart2, bheart3			

3.3 RA prevalence from English national survey data: Health Survey for England (2005) data

An interview with each eligible person was followed by a nurse visit both using computer assisted interviewing. The 2005 survey for adults focused on the health of older people. All adults were asked modules of questions on general health, alcohol consumption, smoking, fruit and vegetable consumption and complementary and alternative medicine. Older informants were also asked about use of health, dental and social care services, cardiovascular disease (CVD), chronic diseases and quality of care, disabilities and falls. Older informants in the boost sample were asked a slightly shorter questionnaire, omitting questions about fruit and vegetable consumption and complementary and alternative medicines (ref. HSfE docs).

3.3.1 HSfE RA outcome variable

The question below is from the HSfE 'Quality of care' section and the instructions were to ask these questions: 'ask all aged 65 and over' (**Figure 30**). Ilsm1-6 variables code for Arthritis/rheumatism/ fibrositis' (value 34). However, since this question does not distinguish between the types of arthritis, it was not used. As for the ELSA data we have shown all the disease categories to demonstrate the breadth of responses available. For other Wave 0 questions (1999 and 2001) we have shown only those relevant to MSK disease. Of course all variables include negative values i.e. Value = -9 Label = Refusal, Value = -8 Label = Don't know, Value = -2 Label = Schedule not applicable, Value = -1 Label = Not applicable.

Pos. = 1059 **Variable =** illsm1-6 **Variable label =** Type of illness - 1st

Value = 1 Label = Cancer (neoplasm) including lumps, masses, tumours and growt
Value = 2 Label = Diabetes. Incl. Hyperglycemia
Value = 3 Label = Other endocrine/metabolic
Value = 4 Label = Mental illness/anxiety/depression/nerves (nes)
Value = 5 Label = Mental handicap
Value = 6 Label = Epilepsy/fits/convulsions
Value = 7 Label = Migraine/headaches
Value = 8 Label = Other problems of nervous system
Value = 9 Label = Cataract/poor eye sight/blindness
Value = 10 Label = Other eye complaints
Value = 11 Label = Poor hearing/deafness
Value = 12 Label = Tinnitus/noises in the ear
Value = 13 Label = Meniere's disease/ear complaints causing balance problems
Value = 14 Label = Other ear complaints
Value = 15 Label = Stroke/cerebral haemorrhage/cerebral thrombosis
Value = 16 Label = Heart attack/angina
Value = 17 Label = Hypertension/high blood pressure/blood pressure (nes)

[7] Not available in ELSA dataset
[8] Not available in ELSA dataset
[9] Not available in ELSA dataset

Value = 18 Label = Other heart problems
Value = 19 Label = Piles/haemorrhoids incl. Varicose Veins in anus.
Value = 20 Label = Varicose veins/phlebitis in lower extremities
Value = 21 Label = Other blood vessels/embolic
Value = 22 Label = Bronchitis/emphysema
Value = 23 Label = Asthma
Value = 24 Label = Hayfever
Value = 25 Label = Other respiratory complaints
Value = 26 Label = Stomach ulcer/ulcer (nes)/abdominal hernia/rupture
Value = 27 Label = Other digestive complaints (stomach, liver, pancreas, bile d
Value = 28 Label = Complaints of bowel/colon (large intestine, caecum, bowel, c
Value = 29 Label = Complaints of teeth/mouth/tongue
Value = 30 Label = Kidney complaints
Value = 31 Label = Urinary tract infection
Value = 32 Label = Other bladder problems/incontinence
Value = 33 Label = Reproductive system disorders
Value = 34 Label = Arthritis/rheumatism/fibrositis
Value = 35 Label = Back problems/slipped disc/spine/neck
Value = 36 Label = Other problems of bones/joints/muscles
Value = 37 Label = Infectious and parasitic disease
Value = 38 Label = Disorders of blood and blood forming organs
Value = 39 Label = Skin complaints
Value = 40 Label = Other complaints
Value = 41 Label = Unclassifiable (no other codable complaint)
Value = 42 Label = Complaint no longer present NB Only use this code if it is a

This variable question (Hediab03) does not distinguish between the types of arthritis. 1,712 (12.88%) respondents (out of 4,269 respondents that were asked this question) answered 'yes', while 2,554 answered 'no'. Only respondents that answered 'yes' were asked further questions (e.g. the type of arthritis).

Pos. = 1671 **Variable =** hediab03 **Variable label =** Doctor diagnosed: Arthritis (including osteoarthritis or rheumatism)
Value = 0 Label = No
Value = 1 Label = Yes

The HSfE heart2 question was used to identify RA cases amongst the respondents that indicated having doctor diagnosed arthritis. 353 respondents indicated having RA, 1,102 indicated not having RA, while 256 didn't know and 1 refused to answer this question.

Pos. = 1682 **Variable =** heart2 **Variable label =** Type of arthritis: Rheumatoid arthritis

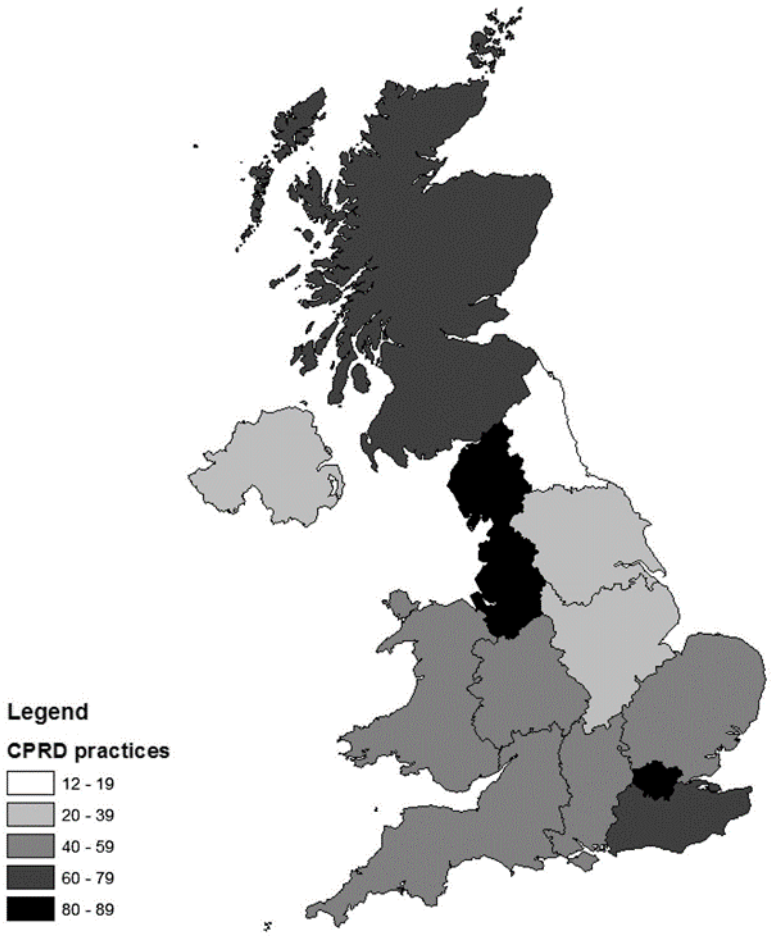
Variable ra was generated to capture the presence/absence of RA. RA prevalence was stratified by age and sex.

- 0 was given in no RA was reported
- 1 was given if RA was reported
- . was given if it was missing (also for respondents that were not asked this question)

3.4 RA prevalence from UK primary care data: Clinical Practice Research Datalink

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



We also used data extracted from the Clinical Practice Research Datalink (http://www.cprd.com/intro.asp) to analyse the prevalence of RA. We identified cases of RA in four ways:

1. cases diagnosed by a doctor
2. cases with linked Hospital Episode Statistics (HES) inpatient diagnosis of RA, which has been validated for other diseases
3. cases which can be inferred from records of symptoms and test results, using the 2010 American College of Rheumatology/European League Against Rheumatism classification criteria for RA[26] even if they have not been explicitly diagnosed by a doctor
4. patients on DMARDs without any other inflammatory arthritis diagnosis and have who have attended a Rheumatology Outpatient Clinic at least twice one year before or after the DMARDs prescription date.

We compiled a comprehensive list of Medcodes for doctor diagnosis of RA, for the symptoms and tests which make up the RA classification and for the disease modifying anti-rheumatic drugs (DMARDs) used on RA patients. To determine the extent of undiagnosed (but diagnosable) RA we then developed a diagnostic algorithm using the 2010 ACR/EuLAR criteria shown in Table 7.

The algorithm itself is also in four parts, including joint involvement, serology, acute phase reactants and duration of symptoms. The medcodes in Table 8 have been divided into those relevant to this case definition and into a fourth section – those relevant to doctor diagnosis of RA. Four data files were derived from the combined data file, one containing records with medcodes in each of these four

divisions. In order to carry out this process in an efficient way these data files were derived from the master file using Perl scripts. The numbers of records and patients in these four files are shown in Table 9. The main RA CPRD extraction took place on 23/01/2015 with additional extraction for a small number of codes on 30/01/2015.

Table 7: The 2010 American College of Rheumatology/European League Against Rheumatism classification criteria for RA

Criteria	Score
Target population (who should be tested?): patients who: **1) have at least one joint with definite clinical synovitis (swelling)*** **2) with the synovitis not better explained by another disease**†	
Classification criteria for RA (score-based algorithm: add score of categories A–D a score of ≥6/10 is needed for classification of a patient as having definite RA)‡	
A. Joint involvement§	
1 large joint¶	0
2–10 large joints	1
1–3 small joints (with or without involvement of large joints)**	2
4–10 small joints (with or without involvement of large joints)	3
>10 joints (at least one small joint)††	5
B. Serology (at least 1 test result is needed for classification)‡‡	
Negative RF *and* negative ACPA	0
Low-positive RF *or* low-positive ACPA	2
High-positive RF *or* high-positive ACPA	3
C. Acute-phase reactants (at least one test result is needed for classification)§§	
Normal CRP *and* normal ESR 0	0
Abnormal CRP *or* normal ESR 1	1
D. Duration of symptoms¶¶	
<6 weeks	0
≥6 weeks	1

***** The criteria are aimed at classification of newly presenting patients. In addition, patients with erosive disease typical of rheumatoid arthritis (RA) with a history compatible with prior fulfilment of the 2010 criteria should be classified as having RA. Patients with long-standing disease, including those whose disease is inactive (with or without treatment) who, based on retrospectively available data, have previously fulfilled the 2010 criteria should be classified as having RA.

† Differential diagnoses differ in patients with different presentations, but may include conditions such as systemic lupus erythematosus, psoriatic arthritis and gout. If it is unclear about the relevant differential diagnoses to consider, an expert rheumatologist should be consulted.

‡ Although patients with a score of less than 6/10 are not classifiable as having RA, their status can be reassessed and the criteria might be fulfilled cumulatively over time.

§ Joint involvement refers to any swollen or tender joint on examination, which may be confirmed by imaging evidence of synovitis. Distal interphalangeal joints, first carpometacarpal joints and first metatarsophalangeal joints are excluded from assessment. Categories of joint distribution are classified according to the location and number of involved joints, with placement into the highest category possible based on the pattern of joint involvement.

¶ 'Large joints' refers to shoulders, elbows, hips, knees and ankles.

** 'Small joints' refers to the metacarpophalangeal joints, proximal interphalangeal joints, second to fifth metatarsophalangeal joints, thumb interphalangeal joints and wrists.

†† In this category, at least one of the involved joints must be a small joint; the other joints can include any combination of large and additional small joints, as well as other joints not specifically listed elsewhere (eg, temporomandibular, acromioclavicular, sternoclavicular, etc.).

‡‡ Negative refers to international unit (IU) values that are less than or equal to the upper limit of normal (ULN) for the laboratory and assay; low-positive refers to IU values that are higher than the ULN but three of less times the ULN for the laboratory and assay; high-positive refers to IU values that are more than three times the ULN for the laboratory and assay. When rheumatoid factor (RF) information is only available as positive or negative, a positive result should be scored as low-positive for RF.

Table 8: Medcodes relevant to the diagnosis of RA[10]

Code type	Text	Medcode
Doctor diagnoses	Rheumatoid arthrit. monitoring	17412
	Polyneuropathy in rheumatoid arthritis	62401
	Myopathy due to rheumatoid arthritis	31209
	Rheumatoid myocarditis	49787
	Rheumatoid carditis	43816
	Rheumatoid lung	9954
	Rheumatoid arthritis and other inflammatory polyarthropathy	27603
	Rheumatoid arthritis	844
	Rheumatoid arthritis of cervical spine	44743
	Other rheumatoid arthritis of spine	44203
	Rheumatoid arthritis of shoulder	21358
	Rheumatoid arthritis of elbow	59738
	Rheumatoid arthritis of distal radio-ulnar joint	63365
	Rheumatoid arthritis of wrist	48832
	Rheumatoid arthritis of MCP joint	42299
	Rheumatoid arthritis of PIP joint of finger	41941
	Rheumatoid arthritis of DIP joint of finger	63198
	Rheumatoid arthritis of hip	49067
	Rheumatoid arthritis of knee	50863
	Rheumatoid arthritis of ankle	51239
	Rheumatoid arthritis of subtalar joint	73619
	Rheumatoid arthritis of talonavicular joint	70658
	Rheumatoid arthritis of other tarsal joint	71784
	Rheumatoid arthritis of 1st MTP joint	51238
	Rheumatoid vasculitis	30548
	Seronegative rheumatoid arthritis	6916
	Rheumatoid bursitis	18155
	Rheumatoid nodule	53621
	Rheumatoid arthritis - multiple joint	31054
	Flare of rheumatoid arthritis	8350
	Felty's syndrome	23552
	Other rheumatoid arthropathy + visceral/systemic involvement	49227

[10] Codes are classified as (1) doctor diagnoses, (2) joint involvement, (3) serology tests and (4) acute phase reactant.

Code type	Text	Medcode
	Rheumatoid lung disease	46436
	Rheumatoid nodule	5723
	Rheumatoid arthropathy + visceral/systemic involvement NOS	37431
	Seropositive errosive rheumatoid arthritis	9707
	Seropositive rheumatoid arthritis, unspecified	12019
	Rheumatoid lung	31724
	Caplan's syndrome	56838
	Fibrosing alveolitis associated with rheumatoid arthritis	28853
	[X]Other seropositive rheumatoid arthritis	93715
	[X]Other specified rheumatoid arthritis	70221
	[X]Seropositive rheumatoid arthritis, unspecified	56202
	Rheumatol. disorder monitoring	29339
	Juvenile rheumatoid arthritis	31360
	O/E - joint swelling	1233
	O/E - swelling - joint	6892
	O/E - joint effusion present	6187
	O/E - joint swelling NOS	22927
	Swelling of joint - effusion	7404
	Joint effusion of unspecified site	1441
	Joint effusion of the shoulder region	21524
	Joint effusion of the upper arm	29700
	Elbow joint effusion	4228
	Wrist joint effusion	56187
	Joint effusion of the hand	15570
	Hip joint effusion	27394
	Knee joint effusion	443
Joint involvement	Joint effusion of the ankle and foot	25934
	Ankle joint effusion	14817
	Effusion of shoulder	24997
	Effusion of elbow	17709
	Effusion of distal radio-ulnar joint	94983
	Effusion of wrist	48812
	Effusion of MCP joint	48127
	Effusion of PIP joint of finger	37131
	Effusion of DIP joint - finger	38980
	Effusion of hip	53659
	Effusion of knee	17658
	Effusion of tibio-fibular joint	65998
	Effusion of ankle	27746
	Effusion of subtalar joint	94322
	Effusion of talonavicular joint	91298
	Effusion of lesser MTP joint	73723
	Effusion of IP joint of toe	62465
	Synovitis of hip	2695
	Synovitis of knee	43238
	Synovitis of elbow	57379
	Synovitis of shoulder	60024
	Shoulder synovitis	16166

Code type	Text	Medcode
	Synovitis of knee	11569
	Synovitis of elbow	17001
	Rheumatoid arthritis of sternoclavicular joint	107963
	Rheumatoid arthritis of acromioclavicular joint	100914
	Rheumatoid arthritis of sacro-iliac joint	100776
	Rheumatoid arthritis of tibio-fibular joint	107791
	Rheumatoid arthritis of lesser MTP joint	99414
	Rheumatoid arthritis of IP joint of toe	107112
	Effusion of joint	479
	Joint effusion of the forearm	51500
	Joint effusion of the pelvic region and thigh	68568
	Joint effusion of the lower leg	34014
	Joint effusion of other specified site	47512
	Effusion of 1st MTP joint	33739
	Chronic joint effusion	29396
	Acute joint effusion	3739
	Effusion of joint NOS	37541
	Intermittent joint effusion	33506
	Synovitis and tenosynovitis	1232
	Synovitis or tenosynovitis NOS	615
	Transient synovitis	16984
	Synovitis NOS	35448
Serology	Rheumatoid factor	14192
	Latex test	14191
	Rose Waaler test	14194
	Rheumatoid factor positive	4502
	R.A. latex test	14190
	Rose Waaler test - sheep cells	15706
	Serum rheumatoid antigen level	27118
	Rheumatoid factor screening test	14195
	Serum rheumatoid antibody level	18901
	Rheumatoid factor IgG level	53299
	Rheumatoid factor IgM level	46370
	IgA rheumatoid factor level	59325
	Rheumatoid factor NOS	14193
	Rheumatoid arthritis screen	16480
	RHEUMATOID FACTOR	83054
	RA SCREEN POSITIVE	78486
	ROSE WAALER TEST	78573
	ROSE WAALER TEST POSITIVE	82923
	RAHA TEST	78411
	RAHA TEST POSITIVE	87351
	[V]Screening for rheumatoid arthritis	6766
Acute phase reactants	C reactive protein abnormal	19809
	Erythrocyte sedimentation rate	46
	ESR abnormal	25450
	ESR raised	14924
	Plasma C reactive protein	14066
	Serum C reactive protein level	14068

Code type	Text	Medcode
Other	[X]Rheumatoid arthritis+involvement/other organs or systems	106440

In addition we found several diagnoses which precluded a diagnosis of RA:

- Psoriatic arthritis
- Spondylitis
- Ankylosing spondylitis
- Inflammatory spondylopathies
- Psoriasis spondylitica
- Sacroiliitis
- Spinal enthesopathy
- Arthritis mutilans

The main RA CPRD extraction took place on 23/01/2015 with additional extraction for a small number of codes on 30/01/2015.

Table 9: Numbers of records and numbers of different patients in files extracted from the master data file; each file contains records with a particular set of medcodes – see Table 8.

Subset of records	Number of records	Number of patients
Doctor diagnosis (X)	342,278	89,675
Joint involvement (A)	322,120	237,052
Serology (B)	911,694	568,901
Acute phase reactants (C)	13,770,770	3,218,652
Patients occurring in each of (A) and at least one of (B) and (C)	1,468,441	136,036

In order to be diagnosed with RA it is necessary that there is some joint involvement and at least one positive test result. Thus those patients meeting the RA diagnosis criteria will be a subset of the 136,036 patients which occur in data set (A) and at least one of data sets (B) and (C). In order to identify which of these 136,036 candidate patients meet the criteria for diagnosis we developed with expert advice a diagnostic algorithm.

3.4.1 Joint involvement

The initial plan was to count the number of small and large joints involved using the CPRD medcodes (Table 8) and score each patient using the first part of the RA diagnosis algorithm. This proved to be intractable for a number of reasons:

- In many cases joint involvement was recorded without specifying which joint or joints were involved.
- The laterality of joint involvement – e.g. whether a left knee or right was involved – was not recorded. This made it impossible to tell whether, for example, two "knee joint" codes referred to the same or different knees.
- The number of joints of a given type was not recorded – e.g. a medcode for "finger joint" could mean that anything from 1 to 8 finger joints were involved.

These considerations made it impossible to count accurately the number of large and small joints involved. It was therefore decided to assign a score of 2 points for the "joint involvement" section of

the algorithm to any patient with joint involvement recorded. A small number of medcodes listed as "joint involvement" codes in Table 8 were excluded, see Table 10.

Table 10: Medcodes excluded from joint involvement section of algorithm

Medcode	Read Term	Reason for exclusion
29700	Joint effusion of the upper arm	Not a joint
38980	Effusion of DIP joint - finger	Joint excluded from diagnosis
100776	Rheumatoid arthritis of sacro-iliac joint	Joint excluded from diagnosis
51500	Joint effusion of the forearm	Not a joint
68568	Joint effusion of the pelvic region and thigh	Not a joint
34014	Joint effusion of the lower leg	Not a joint
33739	Effusion of 1st MTP joint	Joint excluded from diagnosis

In respect to tabulation of joint involvement, the medcodes related to joint involvement are shown in Appendix Table 93, classified into Large Joints, Small Joints, one joint (the DIP joint of finger) which is excluded from consideration in the RA algorithm and those codes which refer to joint synovitis / inflammation without specifying the joint(s) involved. The cumulative total numbers of medcodes related to each joint for the 136,036 candidate patients are also shown in Table 93. The cumulative numbers of medcodes for large and small joints are shown in Table 94 and Table 95. A cross tabulation of numbers of large and small joints involved is shown in Appendix Table 96. The scores which joint involvement contributes to the RA algorithm total are shown below and tabulated tabulated in Appendix Table 97.

- Any joint involvement score 2
- 4–10 small joints (with or without involvement of large joints) score 3
- >10 joints (at least one small joint) score 5

3.4.2 Serology and APR tests results

In some cases the medcodes for Rheumatoid Factor (RF) and Acute Phase Reactant (APR) tests (shown in Table 8) indicate a positive test result. In most cases it was necessary to consider more detailed CPRD data on patient test results. This data included information on the units for the test and the normal range for the test. Information on the upper end of the normal range, to be used as the cut-off for positive test results, was missing in about 40% of cases. This information was supplied by taking the mean of the upper end of normal range data which was not missing. The cut-offs for positive test results used were:

- 20 IU/mL low positive for RF
- 60 IU/ml high positive for RF (set at 3 times the "low positive" value)
- 7.17 mg/L positive for CRP test

For eosinophil sedimentation ratio (ESR) the test cut-off is dependent on both age and sex. Missing cut-offs were therefore supplied using the predicted values from a linear model of the non-missing cut-offs in terms of age and sex (with sex-specific slopes). The results of this linear regression are shown in Table 11.

Table 11: Results of linear regression of cut-offs for positive ESR test on age (years) and sex

Model term	Coefficient	SE	p-value	95% CI for coefficient
Intercept	4.746878	.1325182	0.000	(4.487147 – 5.006609)

Model term	Coefficient	SE	p-value	95% CI for coefficient
Sex = female	1.547737	.075483	0.000	(1.399793 – 1.695682)
Slope (males) / year	.0865374	.0010142	0.000	(.0845496 – .0885252)
Slope (females) / year	.1010686	.0006821	0.000	(.0997318 – .1024055)

The test results were scored in accordance with the specification in Table 7
Serology
2 points if any RF / ACPA test low +ve
3 points if any RF / ACPA test high +ve
Acute-phase reactants
1 point if any CRP or ESR test +ve

Finally, if the time between the first and last record of joint involvement was ≥ 6 weeks a point was scored for "duration of symptoms"). To determine the extent of undiagnosed (but diagnosable) RA we then developed a diagnostic algorithm using the 2010 American College of Rheumatology/European League Against Rheumatism classification criteria for RA.[26]

The case definition is in four parts, including joint involvement, serology, acute phase reactants and duration of symptoms. The medcodes can be divided into those relevant to parts A, B and C of this case definition and into a fourth section – those relevant to doctor diagnosis of RA (this four-way classification is indicated by shading. Four data files were derived from the combined data file, one containing records with medcodes in each of these four divisions. In order to carry out this process in an efficient way these data files were derived from the master file using PERL scripts.

In the EULAR/ACR classification criteria patients can score up to five points for any joint involvement (two points for up to three small joints, three points for four to ten small joints and five points for more than ten joints), up to three points for positive RF/ACPA, one point for positive c-reactive protein (CRP) or erythrocyte sedimentation rate (ESR) test and one point if the duration of symptoms is greater than six weeks. In line with the classification criteria we defined those with a score of six or more as "algorithm-diagnosed RA cases".

3.4.3 Patients with HES RA diagnosis

We used HES outpatient data to identify RA cases and used patient ID as the linkage between CPRD and HES dataset to find any additional RA cases that were not recorded in CPRD. The medcode list used to identify RA cases in HES was

3.4.4 Patients on DMARDs without other inflammatory arthritis diagnosis

We have followed an algorithm to identify any probable RA cases on combination or single DMARDs without any other inflammatory arthritis diagnosis and have seen OPD rheumatologists one year before or after DMARDs prescription. The details of the algorithm is shown in Figure 2.

3.4.5 CPRD risk factors

We used the literature review described in the Background to extract CPRD data on risk factors. There were two main reasons why some risk factors from the literature were not used in the final model. Firstly, the data was not available in CPRD. For example, data on educational level, occupational class and socioeconomic status is very poorly recorded. The occupational classification for which Read codes are available is from a 1986 Office for National Statistics classification so is outdated. Physical activity is also poorly recorded, although this is improving because of the dissemination of the GP Physical Activity Questionnaire (GPPAQ),[27] and the capture of GPPAQ data at the time of NHS Health Checks in particular. CPRD links most patients' data to Index of Multiple Deprivation (IMD) data based on postcode. Secondly, to produce local estimates we use "joint distributions"- cross tabulations which distribute data on each risk factor across the data for all other risk factors- of local risk factor data to which we apply the CPRD prevalence estimates for the same distributions. Hence we can only

use in the final regression model variables which are also available locally. This may cause model performance to deteriorate. We evaluated the extent of this by comparing Receiver Operating Characteristic (ROC) curves for the two models.

Risk factor data were extracted by a defined Read code lists. These are created by searching for relevant Read version 2 5-byte codes using either CPRD's own code browser or using the "NHS browser" maintained by the Health & Social Care information Centre (HSCIC). We used the NHS browser to create code lists for ethnicity, BMI, smoking, and alcohol consumption, by searching relevant read terms or going down the hierarchy of relevant read codes. Social class was defined using deprivation codes.

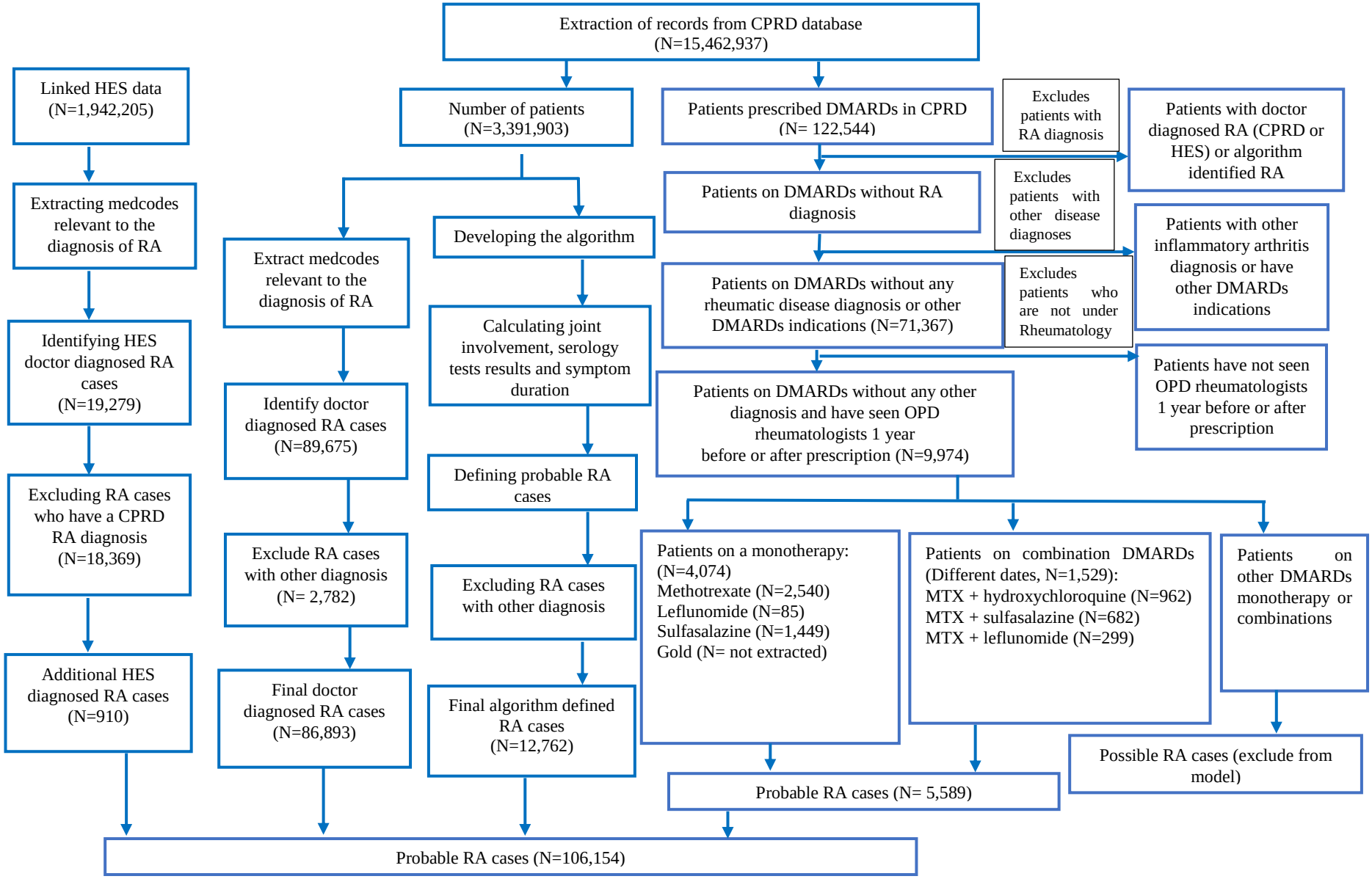
3.4.6 CPRD descriptive analyses

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

3.4.7 CPRD regression modelling

We fitted uni-variate then multivariate logistic regression models for non-specific and radicular back pain as described in previous publications, to produce odds ratios (ORs) and regression coefficients.[28] A range of multivariate regression models were fitted in order to obtain the best performing. We included one additional variable at a time to observe the effects.

Figure 2: Flow chart of the algorithm



3.4.8 Interactions

There is an interaction between the effects of two exposures if the effect of one exposure varies according to the level of the other exposure.[29] For example, there might be an interaction between the back pain risk factors of education level and social class. An alternative term for interaction ifs effect modification. In this example, we can think of this as educational level modifying the effect of social class. The most flexible approach to examine interactions is to use regression models, but when using Mantel-Haenszel methods to control for confounding an alternative is to use a $\chi 2$ test for effect modification, commonly called a test of heterogeneity. Interaction, effect modification and heterogeneity are three different ways of describing the same thing. Log likelihoods are compared in the two models excluding and including the interaction parameters to test the null hypothesis that there is no interaction between selected variables.

We tested for interactions between CPRD predictor variables for risk factors. Interactions were tested between age and gender; BMI and smoking status; BMI and economic activity; economic activity and education; age and education; age and socioeconomic status; age and smoking status; age and economic status; age and BMI.

3.4.9 Internal validation

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

The variables included in the final model are also determined by the availability of local data to match with the model variables. Hence variable selection has to be a compromise between the best model which can be produced from CPRD data and the local variable available.
All statistical analysis was carried out in StataSE14.

3.4.10 External validation

So far we have externally validated the local estimates against the NoAR prevalence data. Given the lack of other similar datasets in the UK we have not been able to carry out other external validations.

3.5 Local prevalence estimates

Derived ORs (or rather, regression coefficients) are used to estimate prevalence in small population subgroups. Local population breakdowns for each risk factor are used, where these are available. ICL has a wide range of small population risk factor prevalence breakdowns, including age, sex, deprivation, smoking, ethnicity, cardiovascular diseases and other disease conditions. The local model uses locally available data. We have developed two methods for producing small population estimates and associated CIs in Stata software. One uses a bootstrapping method to produce repeated samples (Method 1), the other (Method 2) uses sampling-probability weights. Both methods produce CIs for the estimates, which are derived from the variance in the logistic model, not the local populations.

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

When estimating the total effects of individual distal factors on disease, both mediated and direct effects should be considered, because in the presence of mediated effects, controlling for the intermediate factor would attenuate the effects of the more distal one. [30] When estimating the joint effects of the more distal factor and the intermediate one, the mediated and direct effects should be separated, especially if the intermediate factor is affected by other distal factors.

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

3.5.1 Method 1: bootstrapping procedure to produce repeated samples

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

- Use the regression coefficients to generate log odds (since they are from a logistic regression model) for each risk factor subcategory
- Generate a similar table of odds by exponentiation
- Generate a similar table of prevalence in each risk factor subcategory using the epidemiologic formula
- Produce a matching table of small population subcategories. If there are no corresponding local data with a sufficiently granular breakdown e.g. ethnicity by age by sex, this requires deciding how each risk factor should be attributed across other risk factor categories, with evenly as the default. For example, we used the national age/sex/ethnicity breakdown from the Census and age/smoking breakdowns from the HSfE to attribute this data at small population levels. The actual breakdown will be somewhat different and needs to be borne in mind as another source of potential error.
- Multiply the population cells by the corresponding prevalence to estimate the number of people in each cell with the disease

In mathematical notation:

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

where b_0 = regression constant, b_1, b_2, b_3, b_4= other regression coefficients

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

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

Predicted log odds of prevalence in community of n individuals:

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

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

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

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

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

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

Predicted prevalence in community of n individuals:

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

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

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

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

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

In order to find the estimated prevalence for each population, it is necessary to calculate the synthetic prevalence of risk factors for each possible combination of risk factor (as included in the chosen disease-specific logistic regression model). The estimated prevalence for a population is then the weighted average of the prevalence estimates for each combination of risk factors, according to the estimated number of people with each risk factor combination in the population (the population on which synthetic estimates are sought).

These calculations can be carried out in Excel (using VBA code to link prevalence and risk factor spreadsheets with formulae in a workbook) or in Stata software to produce confidence intervals as well as the estimates. We used both methods as a means of validating the synthetic estimation step. The detailed methods of the Stata code we developed and used is included in Annex 1: synthetic estimation using Stata. In summary, within Stata, a new set of variables is created, one for each combination of these risk factors pertinent to the logistic regression model for the chosen disease. With our dataset set up in this way, we can now use Stata's "predict" command to give us the predicted log odds. Then we find the weighted average of these, averaged across all possible combinations of risk factors, using the weights calculated as above (stored in variable named xyz). The

weighted average can be found using the "collapse" command as follows, which results in one line of data per practice or MLSOA (using the population identifier as the by variable) in Stata.

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

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

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

3.5.2 Method 2: Logistic regression and sampling-probability weights

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

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

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

Having estimated the regression model parameters, we used these for out-of-sample prediction of RA prevalence, using the `margprev` add-on Stata package [31 32]. These predicted prevalence estimates were for the sub-populations of patients for 7,692 practices, for 204 clinical care groups (CCGs), and for 6,755 MSOAs, for which information was available on the marginal frequencies of the seven risk factors in the model. We computed estimated prevalences assuming that, within each sub-population, the seven risk factors were mutually statistically independent, implying that we could give each possible combination of the seven risk factors a sampling-probability weight proportional to the product of the proportions of subjects with each of the appropriate risk-factor values. Therefore, for each subpopulation, we had 2x4x5x5x3x4x4=9600 combinations of risk factor values, with proportions of subjects calculated assuming statistical independence, and estimated the expected subpopulation prevalence of RA accordingly. The assumption of statistical independence of risk factors is probably not literally true, but might be expected to give prevalence estimates that are not vastly in error if the effects of the risk factors are not too non-additive. We have not internally or externally validated this method yet.

3.6 Local prevalence estimates for other UK countries

ARUK commissioned us to develop estimates for all UK countries. However it was not possible to develop estimates for Northern Ireland (NI) because of the paucity of local risk factor data. For example, there is no sub-national breakdown of national health survey smoking data, and NI has abandoned collection of practice level smoking data as part of the QOF. The methods used in Scotland and Wales are covered in the Methods section for each of those countries i.e. Section 4.9.1 Methods and Section 4.10.1 Methods respectively.

3.7 Validation of local estimates

In addition to the internal and external validation of the regression models, The local estimates can also be validated by aggregating them to the lowest geography available in the raw data and comparing them, a form of internal validation. These and external validations are shown in the Results. As we have developed and used two methods for producing local estimates and CIs, we could also cross-validate these. However this has proven to be problematic because of the large data volumes in CPRD. We sampled the CPRD database to obtain controls in order to produce a dataset which could be processed in Stata in a reasonable time period. In effect this creates a case/control dataset rather than a population-based or cohort dataset if the whole CPRD population is used. We used this dataset for local estimates with the bootstrapping Stata method, but this produced over-estimates of RA

prevalence. We therefore used the same random sampling ratio for RA cases, but this reduced RA cases to only 3,500, which prevented Stata from fitting the logistic model. We are currently investigating whether the whole CPRD population can be used as controls. In the interim we have used results for Method 2 for the internal and external validations.

4 Results

4.1 RA prevalence from English national survey data: English Longitudinal Study of Ageing

4.1.1 Baseline characteristics of ELSA respondents

Table 12 Characteristics of ELSA respondents'

	RA cases	Non-RA cases	Total
Total number of respondents	2,001 (8.90%)	20,494 (91.10%)	22,495 (100%)
Age (agegrp)			
<44[11]	28 (1.40%)	704 (3.44%)	732 (3.25%)
45-64	1,099 (54.92%)	12,228 (59.67%)	13,327 (59.24%)
65-74	532 (26.59%)	4,463 (21.78%)	4,995 (22.20%)
Over 75	342 (17.09%)	3,099 (15.12%)	3,441 (15.30%)
Gender			
Female	1,257 (62.82%)	11,187 (54.59%)	12,444 (55.32%)
Male	744 (37.18%)	9,307 (45.41%)	10,051 (44.68%)
Ethnicity			
White	1,899 (94.90%)	19,684 (96.05%)	21,583 (95.95%)
Non-white	102 (5.10%)	765 (3.73%)	867 (3.85%)
Missing	0	45 (0.22%)	45 (0.20%)
Education			
NVQ4/NVQ5/Degree or equivalent	127 (6.35%)	2,833 (13.82%)	2,960 (13.16%)
Higher education below degree	196 (9.80%)	2,379 (11.61%)	2,575 (11.45%)
NVQ3/GCE A level equivalent	95 (4.75%)	1,420 (6.93%)	1.515 (6.73%)
NVQ2/GCE O level equivalent	297 (14.84%)	3,275 (15.98%)	3,572 (15.88%)
NVQ1/CSE other grade equivalent	96 (4.80%)	958 (4.67%)	1,054 (4.69%)
Foreign/other	138 (6.90%)	1,532 (7.48%)	1,670 (7.42%)
No qualification	1,033 (51.62%)	7,892 (38.51%)	8,925 (39.68%)
Missing	19 (0.95%)	205 (1%)	224 (1%)
Socioeconomic status			
Higher managerial and professional	79 (3.95%)	1,648 (8.04%)	1,727 (7.68%)
Lower managerial and professional	417 (20.84%)	5,318 (25.95%)	5,735 (25.49%)
Intermediate occupations	123 (6.15%)	1,090 (5.32%)	1,213 (5.39%)
Small employers and own account workers	237 (11.84%)	2,359 (11.51%)	2,596 (11.54%)
Lower supervisory and technical	436 (21.79%)	4,059 (19.81%)	4,495 (19.98%)
Semi-routine occ.	408 (20.39%)	3,131 (15.28%)	3,539 (15.73%)
Routine occ.	231 (11.54%)	1,946 (9.50%)	2,177 (9.68%)
Never worked and long term unemployed	34 (1.70%)	297 (1.45%)	331 (1.47%)
Other	8 (0.40%)	47 (0.23%)	55 (0.24%)
Missing	28 (1.40%)	599 (2.92%)	627 (2.79%)
BMI			
<18.4 underweight	8 (0.40%)	183 (0.89%)	191 (0.85%)

[11] This group will be excluded in further analyses

	RA cases	Non-RA cases	Total
18.5 – 24 normal	258 (12.89%)	3,569 (17.41%)	3,827 (17.01%)
25 – 29 overweight	571 (28.54%)	6,065 (29.59%)	6,636 (29.50%)
>30 obese	908 (45.38%)	8,183 (39.93%)	9,091 (40.41%)
Missing	256 (12.79%)	2,494 (12.17%)	2,750 (12.22%)
Smoking			
Never smoked	165 (8.25%)	3,684 (17.98%)	3,849 (17.11%)
Ex-smoker	1,381 (69.02%)	11,493 (56.08%)	12,874 (57.23%)
Current smoker	435 (21.74%)	3,489 (17.02%)	3,924 (17.44%)
Missing	20 (1.00%)	1,828 (8.92%)	1,848 (8.22%)

4.1.2 RA prevalence in each ELSA wave

Table 13 shows how many RA cases were identified at each ELSA Wave. It is not clear which cases are new and which are old RA cases, so we checked for overlap between them. Table 14 shows unique incident cases at each Wave as overlap between the Waves was checked. If a respondent appeared to be an incident RA case at (e.g. Wave 2) but was also a prevalent RA case at (eg. Wave 1), it would have been considered as an old case at Wave 2 (which was carried forward from Wave 1) and not an incident case at Wave 2. The following were observed and changed accordingly:

- 12 RA cases overlap at Waves 0 and 1. Therefore, number of incident cases at Wave 1 was reduced from 835 to 823.
- No overlap between cases at Waves0- 1 and 2.
- 259 RA cases overlap at Waves 0-2 and 3. Therefore, the number of incident cases at Wave 3 was reduced from 629 to 370
- 301 RA cases overlap between Waves 0-3 and 4. Therefore, the number incident cases at Wave 4 reduced from 704 to 403.
- 375 RA cases overlap between Waves 0-4 and 5. Therefore, the number of incident cases at Wave 5 reduced from 625 to 250.

Table 13 Incident RA cases based on ELSA question

Wave	Incident RA cases at each Wave (based only on question)
Wave 0 (only 2001)	*38 (0.19%)*
Wave 1	835 (6.90%)
Wave 2	117 (1.25%)
Wave 3	629 (6.44%)
Wave 4	704 (6.37%)
Wave 5	625 (6.08%)

Based on this information prevalent RA cases were identified at each Wave Table 14. Variable ra was generated to capture presence/absence of RA:

- 0 was given in no RA was reported in any Waves
- 1 was given if RA was reported at least in on Wave (from Wave 0 2001 to Wave 5)
- . if it was missing

Table 14 RA prevalence and incidence (adjusted for each Wave)

Wave	RA Incidence N (%)	Total # (for incidence)	RA prevalence N (%)	Total # (for prevalence)
Wave 0 (only 2001)	-	-	38 (0.26%)	14,393
Wave 1	823 (7.06%)	11,660	861 (4.82%)	17,845
Wave 2	117 (1.25%)	9,335	978 (5.43%)	18,017
Wave 3	370 (3.79%)	9,771	1,348 (6.84%)	19,719
Wave 4	403 (3.78%)	10,670	1,751 (7.83%)	22,349
Wave 5	250 (2.52%)	9,920	2,001 (8.90%)	22,495

There 2,142 respondents without any information about the absence/presence of RA. Therefore they will be excluded from further analyses. Table 15 shows RA prevalence distribution by sex and age as well as comparing prevalence rates published in the NoAR study (Symmons et al).[33] It is apparent that compared to NoAR ELSA respondents are over-reporting an RA diagnosis.

Table 15 RA prevalence in ELSA stratified by sex and age

Gender	Female				Male				Both sexes			
Age group	<44	45-64	65-74	75+	<44	45-64	65-74	75+	<44	45-64	65-74	75+
Sample size (ELSA)	73	4,478	3,117	4,776	21	3,338	2,925	3,767	94	7,816	6,042	8,543
Number with RA (ELSA)	2	354	327	574	1	189	229	325	3	543	556	899
RA prevalence (ELSA)	2.7%	7.9%	10.5%	12.0%	4.8%	5.7%	7.8%	8.6%	3.2%	7.0%	9.2%	10.5%
Sample size **[33]**	2,516	838	430	401	-	1,201	703	504	-	-	-	-
Number with RA **[33]**	3	14	11	12	-	7	8	11	-	-	-	-
Min RA prevalence (NoAR)[12] **[33]**	0.1%	1.7%	2.6%	3.0%	0.02 [13]	0.6%	1.1%	2.2%	-	-	-	-
RA prevalence (NoAR)[14] **[33]**	0.2%	1.9%	2.8%	3.8%		0.7%	1.3%	2.7%	-	-	-	-
RA prevalence (NoAR)[15] **[33]**	0.2%	2.2%	3.3%	5.4%		0.9%	1.5%	3.1%	-	-	-	-

[12] Prevalence of RA assuming that none of the non-responders to the screening questionnaire and examination had RA. This gives a minimum prevalence but seems the most robust assumption.
[13] Males aged 16–44 yr were not included in the survey. This prevalence figure was calculated by assuming that the female:male ratio of RA in the 16–44 yr age group is the same as that observed in NoAR for the incidence of RA in the same age group (i.e. 6.45:1).
[14] Prevalence of RA assuming that non-responders have same rate of RA as responders.
[15] Prevalence of RA assuming that those who declined examination had the same rate of RA as those who agreed to be examined.

4.1.3 RA incidence and prevalence in ELSA (refined RA case definition; excluded if has hip OA and hip pain)

RA usually affects hands, wrists, elbows, shoulders, knees, ankles and feet. Therefore, if a respondent reported having hip osteoarthritis (OA) (confirmed having OA and hip pain) as well as RA, it is more likely that they only had hip OA (Table 16 and Table 17). Therefore, these respondents were coded as not having RA (n=381). However, it is still possible that they have both hip OA and RA (for example of the hand).

Table 16 Incident hip OA based on ELSA questions

Wave	Incident hip OA cases at each Wave (based on questions only)
Wave 0	-
Wave 1	724 (7.88%)
Wave 2	64 (0.83%)
Wave 3	318 (4.14%)
Wave 4	359 (4.32%)
Wave 5	251 (3.36%)

Table 17 Hip OA incidence and prevalence (adjusted for each wave)

Wave	Hip OA Incidence N (%)	Total # (for incidence)	Hip OA prevalence N (%)	Total # (for prevalence)
Wave 0	-	-	-	-
Wave 1	-	-	723 (7. 88%)	9,184
Wave 2	64 (0.83%)	7,666	787 (7.29%)	10,793
Wave 3	318 (4.14%)	7,690	1,105 (8.74%)	12,645
Wave 4	359 (4.32%)	8,309	1,464 (9.76%)	14,993
Wave 5	251 (3.36%)	7,471	1,715 (11.15%)	15,379

After excluding overlapping cases between RA and hip OA, RA prevalence slightly decreased.

Table 18 RA prevalence stratified by age and sex (excluding overlapping hip OA and RA cases)

Gender	Female				Male				Both sexes			
Age group	**<44**	**45-64**	**65-74**	**75+**	**<44**	**45-64**	**65-74**	**75+**	**<44**	**45-64**	**65-74**	**75+**
Sample size (ELSA)	73	4,478	3,117	4,776	21	3,338	2,925	3,767	94	7,816	6,042	8,543
Number with RA (ELSA)	2	268	245	454	1	162	199	289	3	430	444	743
RA prevalence (ELSA)	2.7%	6.0%	7.9%	9.5%	4.8%	4.9%	6.8%	7.7%	3.2%	5.5%	7.4%	8.7%

4.1.4 RA incidence and prevalence (refined RA case definition, excluded if has hip pain)

In this section, we checked how many RA cases have hip pain as well in each Wave (Table 19). Firstly, prevalent RA cases were compared against the presence/absence of hip pain in a particular Wave (row Wave 5a). However, there was 34.28% of missing hip pain information. To minimise the proportion of missing information, a new variable was created. This variable was given hip pain value from Wave 5, if it was missing – from Wave 4, if that was missing – from Wave 3 etc. In this way, completeness for

this question was maximised and there was only (1.50% missing data) (refer to row Wave 5b). Based on these results, respondents that had RA and hip pain were excluded from being RA case (excluded 454 cases out of 2,001). Therefore, there are 1,547 RA cases and their prevalence was stratified by age and sex (

Table 20 shows ELSA RA prevalence excluding RA cases with hip pain stratified by age and sex. Prevalence is still much higher than that found in NoAR.

Table 20).

Table 19 RA prevalence (with/without hip pain)

Wave	RA prevalence N (%)	Total # (for prevalence)	RA with hip pain prevalence (%)[16]	RA without hip pain prevalence (%)[17]	RA with missing hip pain information prevalence (%)
Wave 0 (2001)	38 (0.26%)	14,393	-	-	
Wave 1	861 (4.82%)	17,845	250 (29.04%)	558 (64.81%)	53 (6.16%)
Wave 2	978 (5.43%)	18,017	177 (18.10%)	532 (54.40%)	269 (27.51%)
Wave 3	1,348 (6.84%)	19,719	218 (16.17%)	765 (56.75%)	365 (27.08%)
Wave 4	1,751 (7.83%)	22,349	294 (16.79%)	902 (51.51%)	555 (31.70%)
Wave 5 a	2,001 (8.90%)	22,495	307 (15.34%)	1,008 (50.37%)	686 (34.28%)
Wave 5 b	2,001 (8.90%)	22,495	454 (22.69%)	1,517 (75.81%)	30 (1.50%)

Table 20 shows ELSA RA prevalence excluding RA cases with hip pain stratified by age and sex. Prevalence is still much higher than that found in NoAR.

Table 20 RA prevalence (excluding RA cases with hip pain) stratified by age and sex

Gender	Female				Male				Both sexes			
Age group	**<44**	**45-64**	**65-74**	**75+**	**<44**	**45-64**	**65-74**	**75+**	**<44**	**45-64**	**65-74**	**75+**
Sample size (ELSA)	73	4,478	3,117	4,776	21	3,338	2,925	3,767	94	7,816	6,042	8,543
Number with RA (ELSA)	2	250	235	453	1	151	181	274	3	401	416	727
RA prevalence (ELSA)	2.7%	5.6%	7.5%	9.5%	4.8%	4.5%	6.2%	7.3%	3.2%	5.1%	6.9%	8.5%

4.1.5 RA incidence and prevalence (refined RA case definition, excluded if has hip pain OR hip replacement due to arthritis)

In this section, we check how many RA cases have hip pain or hip replacements because of arthritis as well in each Wave. Respondents that had RA and hip pain or hip replacement were excluded from being RA case (excluded 481 cases out of 2,001). Therefore there are 1,520 RA cases, and their prevalence was stratified by age and sex (Table 21).

[16] Percentage is obtained using the whole sample (missing information is included as well). For example, out of 861 RA cases at Wave 1: only 808 have information about hip pain, 53 respondents do not have any information about hip pain but are included in calculating percentages.

[17] Same as above. Percentage is obtained using the whole sample (missing information is included as well).

Table 21 RA prevalence (excluding respondents with hip pain or hip replacements because of arthritis) stratified by age and sex

Gender	Female				Male				Both sexes			
Age group	**<44**	**45-64**	**65-74**	**75+**	**<44**	**45-64**	**65-74**	**75+**	**<44**	**45-64**	**65-74**	**75+**
Sample size (ELSA)	73	4,478	3,117	4,776	21	3,338	2,925	3,767	94	7,816	6,042	8,543
Number with RA (ELSA)	2	248	229	440	1	150	180	270	3	398	409	710
RA prevalence (ELSA)	2.7%	5.5%	7.4%	9.2%	4.8%	4.5%	6.2%	7.2%	3.2%	5.1%	6.8%	8.3%

In summary, by comparison with populations surveys with clinical diagnoses, and despite maximal use of other responses in the dataset, it appears that ELSA still over-estimates RA prevalence compared to the NoAR gold standard.

4.2 RA prevalence from English national survey data using Health Survey for England (2005)

RA prevalence was stratified by age and sex (**Table 23**).

Table 22 RA outcome (based on patient reports)

Whether has RA	Frequency	Percentage
No	3,656[18]	85.70%
Yes	353	8.27%
Don't know	256	6.00%
Not answered	1	0.02%
Missing	9,031	-

Table 23 shows RA prevalence stratified by age and sex based on HSfE patient-reported RA, again compared with NoAR data. As with ELSA data, it appears that there is considerable over-reporting of RA diagnoses.

Table 23 RA prevalence stratified by age and sex (based on patient-reported RA)

Gender	Female				Male				Both sexes			
Age group	**<44**	**45-64**	**65-74**	**75+**	**<44**	**45-64**	**65-74**	**75+**	**<44**	**45-64**	**65-74**	**75+**
Sample size (HSfE 2005)	-	-	1,240	1,129	-	-	1,117	780	-	-	2,357	1,909
Number with RA (HSfE 2005)	-	-	101	119	-	-	77	56	-	-	178	175
RA prevalence (HSfE 2005)	-	-	8.2%	10.5%	-	-	6.9%	7.2%	-	-	7.6%	9.2%
Sample size (NoAR) **[33]**	2,516	838	430	401	-	1,201	703	504	-	-	-	-
Number with RA (NoAR) **[33]**	3	14	11	12	-	7	8	11	-	-	-	-

[18] This was obtained by adding respondents that answered 'no' to hediab03 (n=2,554) and heart2 (n=1,102)

Gender	Female					Male			Both sexes			
Min RA prevalence (NoAR) [19] **[33]**	0.1%	1.7%	2.6%	3.0%	0.02 [20]	0.6%	1.1%	2.2%	-	-	-	-
RA prevalence[21] **[33]**	0.2%	1.9%	2.8%	3.8%		0.7%	1.3%	2.7%	-	-	-	-
Ra prevalence[22] **[33]**	0.2%	2.2%	3.3%	5.4%		0.9%	1.5%	3.1%	-	-	-	-

4.2.1 RA prevalence (based on rheumatic disease medication)

HSfE 2005 included a section where a nurse asked about prescribed medicines (Figure 31). Coding for prescribed medicines could be found in HSfE file 5675interviewingdocs.pdf. Table 24 shows the list of RA drugs and their coding. Based on this information, variable 'drugs' was created and it was coded:

- 0 – Not taking RA drugs (if medcnjd was equal to 2; meaning that a respondent indicated not taking any medicines, pills, syrups, ointments, puff OR medcnjd was equal to 1 as it indicated that a person was taking some drugs, but it was excluded from this category if the drug was 'RA drug')
- 1 – Taking RA drugs (if medbi01-22 was equal to 100101 **OR** 100102 **OR** 100103)
- Missing (if not applicable)

Out of 8,774 respondents 8,310 (94.71%) did not take rheumatic disease medication, while 464 (5.29% took RA drugs). Prescribed medicines question was only asked at Nurse visit, therefore, there are only 8,774 recorded answers (Table 25).

Table 24 List of drugs used for RA (based on British National Formulary No. 48 Sept '04)

Medication name	Code
Aspirin	
Analgesic	04.07.01
Antiplatelet	02.09.00
Migraine	04.07.04
Myocardial infarction	02.10.01
Rheumatic disease	10.01.01
Azathioprine	
Myasthenia gravis	10.02.01
Rheumatic disease	10.01.03
Transplant rejection	08.02.01
Ulcerative colitis	01.05.00
Diclofenac sodium	
Eye	11.08.02
Gout (acute attack)	10.01.04
Postoperative pain	15.01.04
Rheumatic disease	10.01.01

[19] Prevalence of RA assuming that none of the non-responders to the screening questionnaire and examination had RA. This gives a minimum prevalence but seems the most robust assumption.
[20] Males aged 16–44 yr were not included in the survey. This prevalence figure was calculated by assuming that the female:male ratio of RA in the 16–44 yr age group is the same as that observed in NoAR for the incidence of RA in the same age group (i.e. 6.45:1).
[21] Prevalence of RA assuming that non-responders have same rate of RA as responders.
[22] Prevalence of RA assuming that those who declined examination had the same rate of RA as those who agreed to be examined.

Medication name	Code
Ureteric coli	07.04.03
Ibuprofen	
Analgesic	04.07.01
Rheumatic disease and gout	10.01.01
Topical antirheumatic	10.03.02
Indometacin (was Indomethacin)	
Gout (acute attack)	10.01.04
Rheumatic disease	10.01.01
Obstetrics	07.01.01
Methotrexate	
Malignant diseases	08.01.03
Rheumatic diseases	10.01.03
Skin (psoriasis)	13.05.03
Naproxen	
Gout (acute attack)	10.01.04
Pain	10.01.01
Rheumatic disease	10.01.01
Prednisolone	
Asthma	03.02.00
Crohn's disease	01.05.00
Eye	11.04.01
Glucocorticoid therapy	06.03.02
Malignant disease	08.02.02
Rectal	01.05.00
Rheumatic disease	10.01.02
Salazopyrin	
Chronic diarrhoea	01.05.00
Rheumatic disease	10.01.03
Voltarol	
Emulgel	10.03.02
Ophtha	11.08.02
Rheumatic disease and gout	10.01.01

Table 25 Distribution of respondents taking RA drugs (broader definition) stratified by age and sex

Gender	Female				Male				Both sexes			
Age group	<44	45-64	65-74	75+	<44	45-64	65-74	75+	<44	45-64	65-74	75+
Sample size (HSfE 2005)	1,985	1,076	935	766	1,644	924	861	583	3,629	2,000	1,620	1,349
Number taking RA drugs (broader def.) (HSfE 2005)	33	75	97	72	16	45	79	47	49	120	176	119
Taking RA drugs (broader def.) (%) (HSfE 2005)	1.7%	7.0%	10.4%	9.4%	1.0%	4.9%	9.2%	8.1%	1.4%	6.0%	9.8%	8.8%

4.2.2 RA prevalence (based on rheumatic disease medication

HSfE 2005 had a section where a nurse asked about prescribed medicines (**Figure 31**). Coding for prescribed medicines can be found below. **Table 26** shows the list of RA drugs and their coding. Based on this information, variable '**drugsben**' was created and it was coded:

- 0 – Not taking RA drugs (if medcnjd was equal to 2; meaning that a respondent indicated not taking any medicines, pills, syrups, ointments, puff OR medcnjd was equal to 1 as it indicated that a person was taking some drugs, but it was excluded from this category if the drug was 'RA drug')
- 1 – Taking RA drugs (if medbi01-22 was equal to 100103)
- Missing (if not applicable)

Out of 8,774 respondents 8,724 (99.43%) did not take rheumatic disease medication, while 50 (0.57% took RA drugs). Prescribed medicines question was only asked at Nurse visit, therefore, there are only 8,774 recorded answers (**Table 27**).

Table 26 List of RA drugs

Medication name	Code
Azathioprine	
Myasthenia gravis	10.02.01
Rheumatic disease	10.01.03
Transplant rejection	08.02.01
Ulcerative colitis	01.05.00
Methotrexate	
Malignant diseases	08.01.03
Rheumatic diseases	10.01.03
Skin (psoriasis)	13.05.03
Salazopyrin	
Chronic diarrhoea	01.05.00
Rheumatic disease	10.01.03

Table 27 Distribution of respondents taking RA drugs stratified by age and sex

Gender	Female				Male				Both sexes			
Age group	<44	45-64	65-74	75+	<44	45-64	65-74	75+	<44	45-64	65-74	75+
Sample size (HSfE 2005)	1,985	1,076	935	766	1,644	924	861	583	3,629	2,000	1,796	1,349
Number taking RA drugs – short list (HSfE 2005)	8	8	14	6	0	6	5	3	8	14	19	9
Taking RA drugs – short list (%) (HSfE 2005)	0.4%	0.7%	1.5%	0.8%	0%	0.7%	0.6%	0.5%	0.2%	0.7%	1.1%	0.7%

4.2.3 RA prevalence (based on rheumatic disease medication and patient-reported RA)

In this section patient-reported RA is combined with the question about RA drugs. A variable **radrugs** is created:

- 0 if does not have RA (based on ra variable; ra=0)
- 1 if reported having RA and taking RA drugs (ra=1 and drugs=1)

- Missing

Out of 4,266 respondents only 62 (1.45%) reported having RA and taking RA drugs (**Table 28**).

Table 28 HSfE RA prevalence based on patient reported RA and use of RA drugs. stratified by age and sex

Gender	Female				Male				Both sexes			
Age group	<44	45-64	65-74	75+	<44	45-64	65-74	75+	<44	45-64	65-74	75+
Sample size (HSfE 2005)	-	-	1,240	1,129	-	-	1,117	780	-	-	2,357	1,909
Number taking RA drugs with reported RA (HSfE 2005)	-	-	25	13	-	-	17	7	-	-	42	20
Taking RA drugs with reported RA (%) (HSfE 2005)	-	-	2.0%	1.2%	-	-	1.5%	0.9%	-	-	1.8%	1.1%

Table 29 presents the overlap between patient-reported RA and respondents that take RA drugs.

Table 29 Overlap between patient-reported RA and RA drugs (broader drug definition)

Whether has RA	Taking drugs for rheumatic disease			
	Not taking	Taking	Missing	Total
Not answered	1	0	0	1
Don't know	21	26	209	256
No	493	207	2,956	3,656
Yes	24	62	267	353
Missing	3,534	169	5,328	9,031
Total	4,073	464	8,760	13,297

4.3 Comparing prevalence obtained using ELSA and HSfE 2005

Table 30 shows different RA prevalence stratified by age and sex that were obtained using three different RA case definitions using both ELSA and HSfE 2005 data sources, and again compared with NoAR data. Both the HSfE and ELSA definitions which rely on patient reports greatly overestimate prevalence compared to NoAR, so they cannot be used in isolation. The HSfE definition "taking RA drugs (broader definition) with patient-reported RA" appears to give similar prevalence to NoAR, and might be more reliable.

Table 30 RA prevalence with different RA definitions (ELSA and HSfE 2005)

Gender	Female				Male				Both sexes				ROC Area[23]
Age group	<44	45-64	65-74	75+	<44	45-64	65-74	75+	<44	45-64	65-74	75+	
RA prev. (ELSA)	2.7%	7.9%	10.5%	12.0%	4.8%	5.7%	7.8%	8.6%	3.2%	7.0%	9.2%	10.5%	0.65

[23] ROC curves for models using different RA definitions could be found in the Appendix 6.3

Gender	Female				Male				Both sexes				ROC Area[23]
Age group	<44	45-64	65-74	75+	<44	45-64	65-74	75+	<44	45-64	65-74	75+	
RA prev. (excl. overlap RA and hip OA) (ELSA)	2.7%	6.0%	7.9%	9.5%	4.8%	4.9%	6.8%	7.7%	3.2%	5.5%	7.4%	8.7%	0.62
RA prevalence (excl. RA cases with hip pain) (ELSA)	2.7%	5.6%	7.5%	9.5%	4.8%	4.5%	6.2%	7.3%	3.2%	5.1%	6.9%	8.5%	0.63
RA prevalence (excl. RA cases with hip pain or hip repl (ELSA)	2.7%	5.5%	7.4%	9.2%	4.8%	4.5%	6.2%	7.2%	3.2%	5.1%	6.8%	8.3%	0.62
RA prevalence (patient reported RA) (HSfE 2005)	-	-	8.2%	10.5%	-	-	6.9%	7.2%	-	-	7.6%	9.2%	-
Taking RA drugs (broader def.) (%) (HSfE 2005)	1.7%	7.0%	10.4%	9.4%	1.0%	4.9%	9.2%	8.1%	1.4%	6.0%	9.8%	8.8%	-
Taking RA drugs (ARUK drug list) (%) (HSfE 2005)	0.4%	0.7%	1.5%	0.8%	0%	0.7%	0.6%	0.5%	0.2%	0.7%	1.1%	0.7%	-
Taking RA drugs with reported RA (%) (HSfE 2005)	-	-	2.0%	1.2%	-	-	1.5%	0.9%	-	-	1.8%	1.1%	-
Min RA prevalence (NoAR)[24] **[33]**	0.1%	1.7%	2.6%	3.0%	0.02[25]	0.6%	1.1%	2.2%	-	-	-	-	-
RA prevalence (NoAR)[26] **[33]**	0.2%	1.9%	2.8%	3.8%		0.7%	1.3%	2.7%	-	-	-	-	-
RA prevalence (NoAR)[27] **[33]**	0.2%	2.2%	3.3%	5.4%		0.9%	1.5%	3.1%	-	-	-	-	-

4.4 ELSA risk factor statistical analyses

A more complete document with outputs from all ELSA models is available on request. Analyses were run excluding respondents younger than 44 years, and their results are presented in the following section. Analyses with this age group included were run and their results can be found in the **Appendix (**Table 89 **-** Table 92**).** ORs were quite similar compared to analyses without this age group. The differences were observed for age group itself between these two types of analyses. Univariate analyses were run for each risk factor using logistic regression (excluding <44 age group, Table 31). All covariates were significant risk factors (except BMI group of normal weight).

[24] Prevalence of RA assuming that none of the non-responders to the screening questionnaire and examination had RA. This gives a minimum prevalence but seems the most robust assumption.

[25] Males aged 16–44 yr were not included in the survey. This prevalence figure was calculated by assuming that the female:male ratio of RA in the 16–44 yr age group is the same as that observed in NoAR for the incidence of RA in the same age group (i.e. 6.45:1).

[26] Prevalence of RA assuming that non-responders have same rate of RA as responders.

[27] Prevalence of RA assuming that those who declined examination had the same rate of RA as those who agreed to be examined.

Table 31 Univariate logistic analysis results using ELSA data

Variable	Odds Ratio	95% CI	p-value
Age			
45-64	1.00		
65-74	1.33	[1.19-1.48]	<0.001
75+	1.23	[1.08-1.40]	0.002
Gender			
Male	1.00		
Female	1.41	[1.28-1.55]	<0.001
Ethnicity			
White	1.00		
Non-white	1.38	[1.12-1.71]	0.003
Education			
NVQ4/NVQ5/Degree or equiv	1.00		
Higher ed below degree	1.84	[1.46-2.31]	<0.001
NVQ3/GCE A level equiv	1.49	[1.14-1.96]	0.004
NVQ2/GCE O level equiv	2.02	[1.63-2.51]	<0.001
NVQ1/CSE other grade equiv	2.24	[1.7-2.94]	<0.001
Foreign/other	2.01	[1.57-2.58]	<0.001
No qualification	2.92	[2.42-3.53]	<0.001
Socioeconomic status			
Higher managerial and professional occup	1.00		
Lower managerial and professional occup	1.64	[1.28-2.09]	<0.001
Intermediate occupations	2.35	[1.76-3.15]	<0.001
Small employers and own account workers	2.10	[1.61-2.72]	<0.001
Lower supervisory and technical occup	2.24	[1.75-2.87]	<0.001
Semi-routine occupations	2.72	[2.12-3.48]	<0.001
Routine occupations	2.48	[1.9-3.22]	<0.001
Never worked or long term unemployed	2.39	[1.57-3.64]	<0.001
Other	3.55	[1.62-7.77]	0.002
BMI			
<18.4 underweight			
18.5-24 normal weight	1.65	[0.81-3.39]	0.170
25-29 overweight	2.15	[1.06-4.39]	0.035
>30 obese	2.54	[1.25-5.17]	0.010
Smoking status			
Never smoked			
Ex-smoker	2.68	[2.27-3.17]	<0.001
Current smoker	2.78	[2.31-3.35]	<0.001

Table 32 Multivariate logistic analysis results using ELSA data

Variable	Odds Ratio	95% CI	p-value
Age			
45-64	1.00		
65-74	1.15	[1.01-1.3]	0.030
75+	1.02	[0.88-1.19]	0.777
Gender			
Male	1.00		
Female	1.54	[1.37-1.71]	<0.001
Ethnicity			
White	1.00		
Non-white	1.84	[1.43-2.37]	<0.001
Education			
NVQ4/NVQ5/Degree or equiv	1.00		
Higher ed below degree	1.42	[1.1-1.85]	0.008
NVQ3/GCE A level equiv	1.24	[0.91-1.68]	0.175
NVQ2/GCE O level equiv	1.54	[1.21-1.98]	0.001
NVQ1/CSE other grade equiv	1.72	[1.25-2.37]	0.001
Foreign/other	1.46	[1.1-1.94]	0.009
No qualification	2.00	[1.58-2.53]	<0.001
Socioeconomic status			
Higher managerial and professional occup	1.00		
Lower managerial and professional occup	1.26	[0.96-1.67]	0.099
Intermediate occupations	1.52	[1.08-2.12]	0.015
Small employers and own account workers	1.46	[1.08-1.97]	0.014
Lower supervisory and technical occup	1.45	[1.09-1.93]	0.011
Semi-routine occupations	1.53	[1.15-2.05]	0.004
Routine occupations	1.33	[0.98-1.82]	0.069
Never worked or long term unemployed	1.45	[0.9-2.35]	0.124
Other	2.13	[0.85-5.34]	0.107
BMI			
<18.4 underweight	1.00		
18.5-24 normal weight	2.12	[1.03-4.37]	0.042
25-29 overweight	2.90	[1.42-5.96]	0.004
>30 obese	3.32	[1.62-6.8]	0.001
Smoking status			
Never smoked	1.00		
Ex-smoker	2.60	[2.16-3.14]	<0.001
Current smoker	2.74	[2.23-3.38]	<0.001

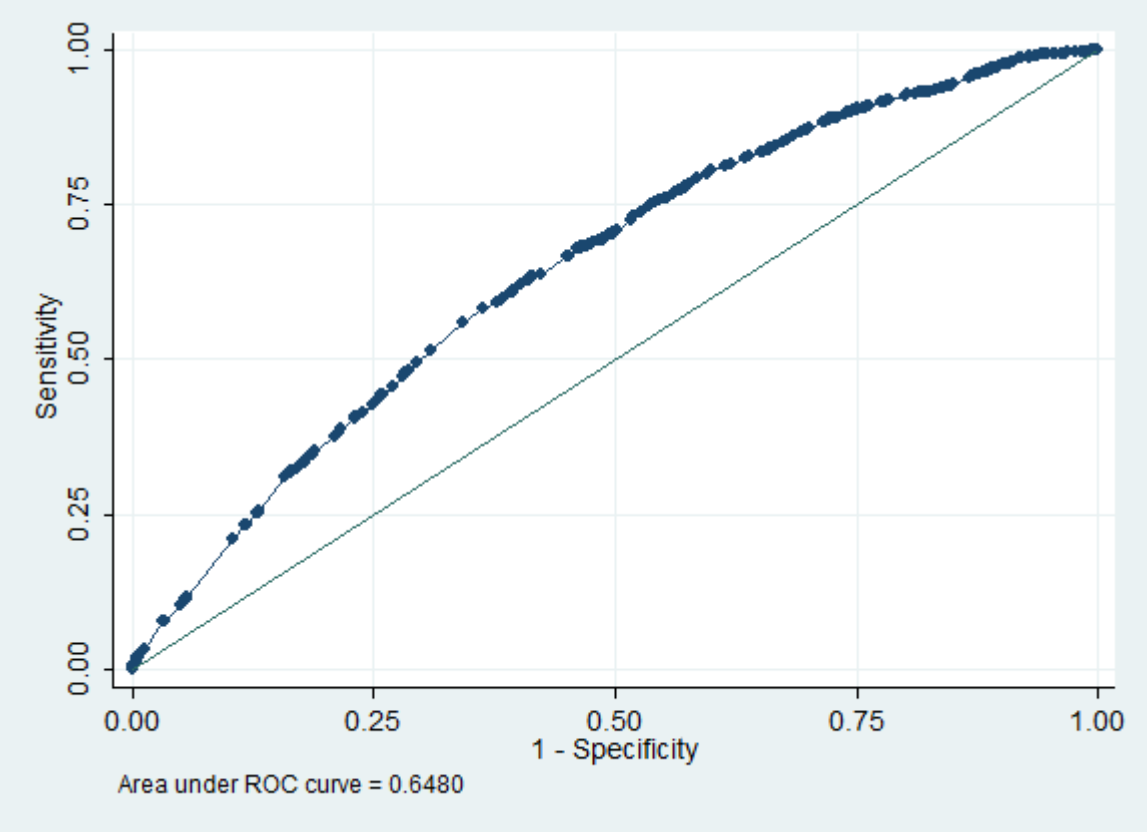
Table 33 Automatic forward stepwise logistic regression analysis using ELSA data

Variable	Odds Ratio	95% CI	p-value
Age			
45-64	1.00		
65-74	1.14	[1.01-1.28]	0.033
75+	1.00		
Gender			
Male	1.00		
Female	1.56	[1.4-1.73]	<0.001
Ethnicity			
White	1.00		
Non-white	1.88	[1.46-2.42]	<0.001
Education			
NVQ4/NVQ5/Degree or equiv	1.00		
Higher ed below degree	1.53	[1.19-1.98]	0.001
NVQ3/GCE A level equiv	1.37	[1.01-1.85]	0.042
NVQ2/GCE O level equiv	1.74	[1.38-2.21]	<0.001
NVQ1/CSE other grade equiv	1.98	[1.46-2.69]	<0.001
Foreign/other	1.66	[1.27-2.18]	<0.001
No qualification	2.32	[1.88-2.87]	<0.001
BMI			
<18.4 underweight	1.00		
18.5-24 normal weight	2.11	[1.02-4.34]	0.044
25-29 overweight	2.88	[1.41-5.91]	0.004
>30 obese	3.31	[1.62-6.78]	0.001
Smoking status			
Never smoked	1.00		
Ex-smoker	2.60	[2.16-3.13]	<0.001
Current smoker	2.76	[2.24-3.4]	<0.001

Multivariate analyses were run using logistic regression. Information from 16,996 respondents' was analysed in this model. The results are presented in the Table 32. Stepwise forward/backward options were added to the logistic function to automatically select significant variables in the model. Forward stepwise regression output is presented in the Table 33, while backward stepwise function output is in the Appendix - Table 35 (since both outputs are quite similar). The socioeconomic status variable ('occup') was dropped from both models. However it should be noted that ELSA data is the least reliable in terms of actual prevalence.

The ROC curve in Figure 3 was produced using an automatic (stepwise) forward logistic regression model**.** The area under the ROC curve was 0.6480 ±0.0068 (95% CI 0.6348-0.6613).

Figure 3 ROC curve for automatic forward stepwise regression model using ELSA data



4.4.1 Internal validation of ELSA: How good is our model at predicting RA caseness?

We could use the ELSA automatic stepwise forward model to predict the probability of individual being RA case in ELSA data set. In **Figure 4** the two box plots show the predicted probability of people with RA caseness among the non-RA and RA groups. Since we have a binary response model, we can choose a cut-off point on the predicted probability to separate the predicted RA cases (with higher predicted probability) from the predicted non-RA 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-RA people being classified as predicted severe RA cases, or RA cases being classified as predicted non-RA cases. Therefore, we use sensitivity and specificity plots to help with this decision.

Figure 4 Predicted probabilities of being RA case

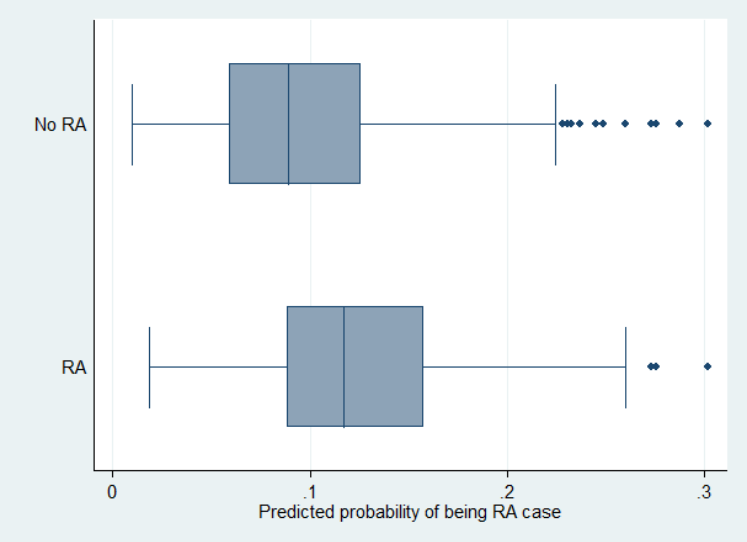
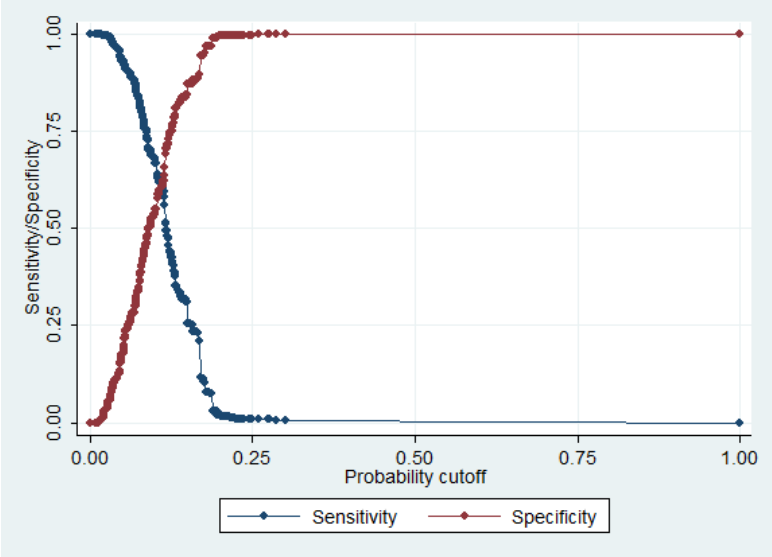


Figure 5 Sensitivity/specificity versus probability cut-off



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. We choose the cut-off probability where sensitivity and specificity lines cross. At cut-off point of predicted probability 0.11, the sensitivity and specificity both reach 60.92% and 60.62%, respectively. Applying this cut-off probability to our data, the following table shows the comparison between predicted and true cases of RA in ELSA (**Table 34**).

Table 34 Predicted RA caseness with different cut-off probabilities

Probability cut-off	0	0.025	0.04	0.05	0.06	0.07	0.1	0.11	0.13	0.15	0.25
Sensitivity (%)	100%	99.40%	96.72%	93.02%	89.86%	87.11%	67.96%	**60.92%**	38.78%	25.54%	0.95%
Specificity	0%	3.39%	11.03%	17.80%	25.62%	30.12%	53.65%	**60.62%**	78.41%	86.92%	99.75%

Probability cut-off	0	0.025	0.04	0.05	0.06	0.07	0.1	0.11	0.13	0.15	0.25
(%)											
True positive	1,676	1,666	1,621	1,559	1,506	1,460	1,139	**1,021**	650	428	16
False positive	0	10	55	117	170	216	537	**655**	1,026	1,248	1,660
True negative	0	519	1,690	2,727	3,925	4,615	8,219	**9,287**	12,013	13,316	15,281
False negative	15,320	14,801	13,630	12,593	11,395	10,705	7,101	**6,033**	3,307	2,004	39

4.4.2 HSfE risk factor statistical analysis

We fitted logistic regression models to HSfE 2005 data in the same way as for ELSA. Table 35 shows a multivariable automatic stepwise backward logistic regression model. For brevity other models are shown in the Appendix Table 89-Table 92, including data from the <44 age group. Because it was obviously not a candidate data source we did not carry out further internal validation.

Table 35: HSfE automatic stepwise backward logistic regression model

Variable	Odds Ratio	95% CI	p-value
Age			
45-64	1.00		
65-74	1.14	[1.01-1.28]	0.032
75+	1.00		
Gender			
Male	1.00		
Female	1.56	[1.4-1.74]	<0.001
Ethnicity			
White	1.00		
Non-white	1.87	[1.45-2.4]	<0.001
Education			
NVQ4/NVQ5/Degree or equiv	1.00		
Higher ed below degree	1.36	[1.09-1.7]	0.007
NVQ3/GCE A level equiv	1.00		
NVQ2/GCE O level equiv	1.54	[1.26-1.88]	<0.001
NVQ1/CSE other grade equiv	1.75	[1.33-2.32]	<0.001
Foreign/other	1.47	[1.16-1.87]	0.002
No qualification	2.05	[1.73-2.43]	<0.001
BMI			
<18.4 underweight	1.00		
18.5-24 normal weight	2.10	[1.02-4.34]	0.044
25-29 overweight	2.89	[1.41-5.92]	0.004
>30 obese	3.32	[1.62-6.79]	0.001
Smoking status			
Never smoked	1.00		
Ex-smoker	2.61	[2.17-3.14]	<0.001
Current smoker	2.78	[2.26-3.43]	<0.001

4.5 CPRD RA definitions, incidence & prevalence

4.5.1 Data extraction

The number of records and the number of different patients in the extractions from each CPRD database are shown in **Table 36**. The six data files were combined to produce a final data file with 15,462,937 records for 3,391,903 patients.

Table 36: Numbers of records and numbers of different patients in the extractions from the CPRD databases.

Extraction	Data extracted from database	Number of records	Number of patients
Initial extraction	Clinical	542,299	247,498
	Referral	368,382	212,630
	Test	14,326,268	3,250,092
Additional extraction	Clinical	213,927	150,541
	Referral	11,675	10,580
	Test	386	254
All files combined		15,462,937	3,391,903

4.5.2 Doctor diagnosed RA cases

Of the 3,391,903 patients in data set (X), 89,675 patients had a doctor diagnosis of RA. Of these we discounted 3,391 patients (3.78%) because they had an alternative diagnosis which precludes an RA diagnosis (See Appendix 6). This left 86,284 cases of doctor diagnosed RA in the CPRD extract.

4.5.3 Algorithm identified "probable RA cases"

The scores for the "joint" section of the algorithm are shown for algorithms A and B in Table 37. The total scores for the algorithms are shown in Table 40. The number of doctor and algorithm diagnosed cases of RA are shown in Table 41, with and without the exclusion of cases with competing diagnoses.

The numbers of additional cases identified by the algorithm are considerably lower than those identified from earlier versions. This is due in part to the additional restriction that APR tests must occur within 3 months of joint involvement in order to count. Mostly, however, this is due to the bug which has now been fixed which (unfortunately) caused the number of algorithm cases to be artificially inflated in earlier analyses.

The very small number of additional cases identified using algorithm A means that if the algorithm is to be used it will be necessary to adopt version B in which the minimum score for joint involvement is 2. This raises legitimate questions of whether these additional "cases" are in fact cases of RA or merely "those at high risk". This will need to be investigated further through a number of validation strategies.

Table 37: Score from joint section of the algorithm for "Algorithm A" and "Algorithm B"

Score from joint section of algorithm	Algorithm A		Algorithm B	
	Frequency	%	Frequency	%
0	1,120	0.82	1,120	0.82
1	133,914	98.44	0	0.00
2	988	0.73	134,902	99.17
3	14	0.01	14	0.01

Table 38: Total score from the algorithm for "Algorithm A" and "Algorithm B"

Total score from algorithm	Algorithm A		Algorithm B	
	Frequency	%	Frequency	%
0	966	0.71	966	0.71
1	51,479	37.84	0	0.00
2	54,438	40.02	51,920	38.17
3	1,222	0.90	54,527	40.08
4	11,346	8.34	701	0.52
5	16,483	12.12	11,418	8.39
6	98	0.07	16,493	12.12
7	4	0.00	11	0.01

Table 39: Number of doctor diagnosed cases of RA and additional cases identified by algorithms A and B[28]

Cases of RA		Before exclusions	After exclusions
Doctor diagnosed cases		89,675	88,299
Additional algorithm diagnosed cases	Algorithm A	70	68
	Algorithm B	13,321	12,928

4.5.4 Additional RA cases from HES outpatient dataset

We aimed to find out any cases with RA diagnosis in HES outpatient data but no RA diagnosis recorded in CPRD. HES outpatient data are linked with CPRD by same patient ID. There are a total of 19,279 patients in HES database that have a diagnosis of RA. In CPRD dataset, we identified a total of 86,879 cases with doctor diagnosed RA, however, the number of RA patients may be underestimated. There are 910 additional RA cases from HES dataset. Among the 910 additional RA cases, there are 13 patients also identified by the algorithm and 101 matched with the patients that were on DMARDs but without any other inflammatory arthritis diagnosis.

4.5.5 Patients on DMARDs without other inflammatory arthritis diagnosis

We dropped patients only on prednisolone and have a total of 122,544 patients on DMARDs. Among these patients, there are 41,830 patients with diagnosed RA. After excluding these RA patients, there are 80,714 patients without an RA diagnosis. We further excluded patients with other inflammatory arthritis disease diagnoses (n=9,347) to leave a total of 71,367 patients on DMARDs without any RA or inflammatory arthritis diagnosis. We used the previous code list to exclude patients with other inflammatory diseases, please see Appendix 1 for details.

4.5.6 CPRD prevalence and incidence

Prevalence and incidence of RA in the CPRD data were calculated for doctor diagnosed RA and for algorithm diagnosed RA (or "high risk of RA") using algorithm B. Prevalence was inferred from cumulative incidence, with RA cases removed only at death.

The prevalence of RA for the years 1960-2014 is shown in Table 40. As noted before, there are considerable numbers of "historical diagnosis" for doctor diagnosed RA but very few historical diagnoses for algorithm diagnosed RA. The prevalence of RA is estimated at around 0.49% for doctor diagnosed RA, rising to 0.58% if algorithm cases are included. The incidence of RA for the years 1960-

[28] Figures are shown for before and after the exclusion of cases with alternative diagnoses which preclude a diagnosis of RA

2014 is shown in Table 41. Table 42 to Table 45 show the prevalence and incidence of doctor diagnosed and algorithm diagnosed RA in the years 2000-2014, broken down by age group and sex.

Table 40: Prevalence of doctor- and algorithm-diagnosed RA in the CPRD data: 1960-2014.

Year	Prevalence of doctor diagnosed RA (cases per million)	Prevalence of additional algorithm diagnosed RA (cases per million)	Prevalence of total RA – doctor and algorithm diagnosed (cases per million)	Percentage change in the doctor diagnosed prevalence from addition of algorithm diagnosed cases
1960	162.2	0	162.2	0
1961	173.3	0	173.3	0
1962	185.7	0	185.7	0
1963	197.2	0	197.2	0
1964	209	0	209	0
1965	226.7	0	226.7	0
1966	240.2	0	240.2	0
1967	254.8	0	254.8	0
1968	268.9	0	268.9	0
1969	280.7	0	280.7	0
1970	304.7	0	304.7	0
1971	322	0	322	0
1972	338.7	0	338.7	0
1973	359	0	359	0
1974	380.4	0	380.4	0
1975	404	0	404	0
1976	425.7	.2	426	.1
1977	451.2	.2	451.4	.1
1978	476.5	.2	476.7	0
1979	499	.2	499.3	0
1980	534.1	.3	534.5	.1
1981	556.3	.4	556.8	.1
1982	584	.5	584.6	.1
1983	611.2	.5	611.7	.1
1984	640.9	.6	641.5	.1
1985	671	.8	671.8	.1
1986	703.7	.9	704.6	.1
1987	740.2	1.3	741.4	.2
1988	785.3	1.3	786.6	.2
1989	858.4	1.8	860.2	.2
1990	983.2	3.6	986.8	.4
1991	1109.4	7	1116.4	.6
1992	1236.5	12.9	1249.4	1
1993	1361.4	24.1	1385.5	1.8
1994	1485.9	36.7	1522.6	2.5
1995	1601.3	49	1650.4	3.1
1996	1722.1	64.4	1786.5	3.7
1997	1841.5	84.2	1925.6	4.6
1998	1957.8	108.6	2066.3	5.5
1999	2076	141.6	2217.6	6.8

Year	Prevalence of doctor diagnosed RA (cases per million)	Prevalence of additional algorithm diagnosed RA (cases per million)	Prevalence of total RA – doctor and algorithm diagnosed (cases per million)	Percentage change in the doctor diagnosed prevalence from addition of algorithm diagnosed cases
2000	2204.6	186.5	2391.2	8.5
2001	2361.3	238.9	2600.2	10.1
2002	2522.4	296.3	2818.6	11.7
2003	2695.8	364.7	3060.5	13.5
2004	2908.8	429.4	3338.2	14.8
2005	3103.1	495.6	3598.7	16
2006	3293.7	556.7	3850.3	16.9
2007	3467.4	613.5	4080.9	17.7
2008	3633.2	676.5	4309.7	18.6
2009	3803.2	730.5	4533.7	19.2
2010	3948.3	783.8	4732.1	19.9
2011	4094.9	830.2	4925.1	20.3
2012	4272.7	874.8	5147.5	20.5
2013	4552.3	911.9	5464.2	20
2014	4877.4	942.1	5819.4	19.3

Table 41: Incidence of doctor- and algorithm-diagnosed RA in the CPRD data: 1960-2014.

Year	Incidence of doctor diagnosed RA (cases per million)	Incidence of additional algorithm diagnosed RA (cases per million)	Incidence of total RA – doctor and algorithm diagnosed (cases per million)	Percentage change in doctor diagnosed incidence from addition of algorithm diagnosed cases
1960	24.2	0	24.2	0
1961	15.9	0	15.9	0
1962	17.8	0	17.8	0
1963	17.5	0	17.5	0
1964	18.5	0	18.5	0
1965	24.4	0	24.4	0
1966	20.3	0	20.3	0
1967	22.7	0	22.7	0
1968	22.1	0	22.1	0
1969	20.4	0	20.4	0
1970	32.2	0	32.2	0
1971	26.7	0	26.7	0
1972	26.8	0	26.8	0
1973	29.4	0	29.4	0
1974	30.8	0	30.8	0
1975	33.4	0	33.4	0
1976	31	.2	31.2	.8
1977	35	0	35	0
1978	35.3	0	35.3	0
1979	33.9	0	33.9	0
1980	46.4	.1	46.5	.2
1981	34.1	.1	34.2	.3

Year	Incidence of doctor diagnosed RA (cases per million)	Incidence of additional algorithm diagnosed RA (cases per million)	Incidence of total RA – doctor and algorithm diagnosed (cases per million)	Percentage change in doctor diagnosed incidence from addition of algorithm diagnosed cases
1982	39.1	.1	39.2	.3
1983	38.6	0	38.6	0
1984	40.2	.1	40.3	.3
1985	40.9	.1	41	.2
1986	45.1	.1	45.2	.2
1987	48.6	.4	49	.8
1988	56.5	.1	56.6	.2
1989	87.7	.5	88.1	.5
1990	147.1	1.8	148.9	1.3
1991	160.7	3.4	164.1	2.1
1992	166.5	6	172.5	3.6
1993	170.3	11.3	181.6	6.6
1994	175.9	12.8	188.7	7.3
1995	171.2	12.6	183.8	7.4
1996	175.4	15.5	190.9	8.8
1997	176.6	20.4	197	11.5
1998	185.3	25	210.3	13.5
1999	187.7	33	220.7	17.6
2000	211.6	45.5	257.1	21.5
2001	241.9	52.9	294.8	21.9
2002	249.9	59.4	309.3	23.8
2003	271	71.3	342.2	26.3
2004	317.8	69.6	387.4	21.9
2005	304.9	69.2	374.1	22.7
2006	295	66.7	361.7	22.6
2007	284	64.3	348.4	22.6
2008	278.3	71.2	349.5	25.6
2009	281.5	63.7	345.2	22.6
2010	257.9	61.4	319.3	23.8
2011	259.6	56.5	316	21.8
2012	279	55	334	19.7
2013	388.4	48.7	437.1	12.5
2014	410.1	41	451.1	10

It is notable from Tables 47-50 that the prevalence / incidence of algorithm cases as a percentage of doctor diagnosed cases decreases with increasing age (see the four columns at the right of each table). This is consistent with seeing the algorithm diagnosed "cases" as an at risk group who fall short of having RA but are likely to go on and develop it.

Table 42: Prevalence of doctor- and algorithm-diagnosed RA in the CPRD data: 2000-2014: males only, broken down by age at diagnosis.

Year	Prevalence of doctor diagnosed RA (cases per million people)				Prevalence of additional algorithm diagnosed RA (cases per million people)				Percentage change in doctor diagnosed prevalence from addition of algorithm diagnosed cases			
	18-44	45-64	65-74	75+	18-44	45-64	65-74	75+	18-44	45-64	65-74	75+
2000	287.3	2486.6	6039.1	7594.3	50.4	346.1	441.6	202.8	17.5	13.9	7.3	2.7
2001	305.7	2582.9	6356.3	7856.3	64.2	429.1	566.8	292.1	21	16.6	8.9	3.7
2002	320.5	2678.6	6684.2	8091.4	76.2	524.1	710.8	390	23.8	19.6	10.6	4.8
2003	331.9	2799.9	7020.9	8435.9	88.6	619.5	858.2	484.2	26.7	22.1	12.2	5.7
2004	344.3	2917.2	7399	8847.7	97.6	714	938.6	614.4	28.3	24.5	12.7	6.9
2005	360.6	3018.1	7612.8	9299.9	109.1	793.4	1109.1	719.3	30.2	26.3	14.6	7.7
2006	384.1	3125.4	7878.3	9539.6	120.4	843.4	1242.4	822.3	31.3	27	15.8	8.6
2007	394.5	3214.4	8079.4	9718.2	132.5	896.4	1360.4	896.9	33.6	27.9	16.8	9.2
2008	404.8	3259.8	8243.2	9948.1	143	958.8	1437	1040.6	35.3	29.4	17.4	10.5
2009	426.8	3249.3	8483.2	10293.5	147.8	984.2	1529.5	1154.9	34.6	30.3	18	11.2
2010	437.8	3251.8	8538.6	10407	159	995.6	1659.5	1291.2	36.3	30.6	19.4	12.4
2011	440.6	3260.7	8622.8	10468.7	161	1004	1758.6	1373.8	36.5	30.8	20.4	13.1
2012	452.3	3241.7	8693	10672.5	159.4	1021.3	1780.6	1456.5	35.2	31.5	20.5	13.6
2013	476.7	3350.9	8991.7	11033	154.3	1013.7	1835.2	1525	32.4	30.3	20.4	13.8
2014	506.3	3465.7	9341.1	11510.3	150.1	990.1	1893	1565.6	29.6	28.6	20.3	13.6

Table 43: Prevalence of doctor- and algorithm-diagnosed RA in the CPRD data: 2000-2014: females only, broken down by age at diagnosis.

Year	Prevalence of doctor diagnosed RA (cases per million people)				Prevalence of additional algorithm diagnosed RA (cases per million people)				Percentage change in the doctor diagnosed prevalence from the addition of algorithm diagnosed cases			
	18-44	45-64	65-74	75+	18-44	45-64	65-74	75+	18-44	45-64	65-74	75+
2000	613.2	5499.2	11472.2	13710.9	100.6	694.9	522.1	248.1	16.4	12.6	4.6	1.8
2001	649.4	5793.9	12096.8	14606.8	126	862.4	717.1	330.4	19.4	14.9	5.9	2.3
2002	686.4	6130.3	12552	15498	153	1038.4	906	435.6	22.3	16.9	7.2	2.8
2003	728	6484.8	13056.2	16336.1	185.3	1270.4	1134.3	560.6	25.4	19.6	8.7	3.4
2004	785.8	6939.8	13953	17237.4	212	1468.7	1382.7	707.8	27	21.2	9.9	4.1
2005	841	7275.6	14857.5	17905.8	238.7	1637.9	1703	827.7	28.4	22.5	11.5	4.6
2006	903.7	7513.9	15688.7	18495.8	259.8	1816	1929.3	946.8	28.7	24.2	12.3	5.1
2007	952.1	7747.1	16442.1	18900.5	272.3	1940.5	2204.8	1083.8	28.6	25	13.4	5.7
2008	1000.6	7865.1	17099.8	19264.7	291.3	2063.8	2441.1	1231.8	29.1	26.2	14.3	6.4
2009	1067.3	7921.1	17500.6	19633.2	306	2154.4	2655.9	1377.9	28.7	27.2	15.2	7
2010	1112.8	7984.9	17816.9	19776.3	314	2220.8	2894.3	1493.3	28.2	27.8	16.2	7.6
2011	1147.5	7966	18183.3	20056	319.7	2250.9	3153	1620.4	27.9	28.3	17.3	8.1
2012	1194.2	7952.9	18616.4	20422.4	328.4	2261.8	3346.6	1761.6	27.5	28.4	18	8.6
2013	1278.4	8168.7	19371.2	20977.4	332.1	2252.9	3503.8	1890.9	26	27.6	18.1	9
2014	1358.4	8451.2	20153.1	21797	333	2205	3663.2	2014.1	24.5	26.1	18.2	9.2

Table 44: Incidence of doctor- and algorithm-diagnosed RA in the CPRD data: 2000-2014: males only, broken down by age at diagnosis

Year	Incidence of doctor diagnosed RA (cases per million people)				Incidence of additional algorithm diagnosed RA (cases per million people)				Percentage change in the doctor diagnosed incidence from the addition of algorithm diagnosed cases			
	18-44	45-64	65-74	75+	18-44	45-64	65-74	75+	18-44	45-64	65-74	75+
2000	27.9	262.1	646.4	598.2	16	87.7	96.4	64.9	57.5	33.5	14.9	10.8
2001	36.7	293.3	663.2	632.3	17.1	98.1	94.7	69.2	46.5	33.4	14.3	10.9
2002	32.5	303.1	761.1	633.5	16.5	111.6	126.6	87.9	50.8	36.8	16.6	13.9
2003	34.5	325.2	793.1	676.5	18.7	108.7	141.2	99.6	54.3	33.4	17.8	14.7
2004	35.5	339.6	909.2	728	15.1	112.8	114.6	110.2	42.6	33.2	12.6	15.1
2005	36.5	351.4	705.9	714.5	17.8	107.8	153.7	75	48.6	30.7	21.8	10.5
2006	45.3	341	721.2	636.9	19.5	86.1	128.4	98.8	43	25.3	17.8	15.5
2007	39.4	304.1	726	566	20.1	92.7	138.2	71.1	51	30.5	19	12.6
2008	38.2	303.7	706.6	552.2	18.6	107.3	134.6	115.2	48.8	35.3	19	20.9
2009	46.3	285	717.7	607.3	12.8	90.7	116.9	82.2	27.7	31.8	16.3	13.5
2010	38.9	280.4	586.7	450.2	19.6	83.8	114.8	79.7	50.5	29.9	19.6	17.7
2011	31	268.3	594.4	502.5	11.9	68.1	111	79.5	38.5	25.4	18.7	15.8
2012	40.7	256.5	656.6	517.2	10.3	67.1	91.6	73.4	25.4	26.2	14	14.2
2013	55.7	372.8	883	714.5	8.6	56.4	97.7	55.6	15.4	15.1	11.1	7.8
2014	60.3	387.6	868.2	609.1	7	36.7	86.2	53.8	11.7	9.5	9.9	8.8

Table 45: Incidence of doctor- and algorithm-diagnosed RA in the CPRD data, 2000-2014: females only, broken down by age at diagnosis.

Year	Incidence of doctor diagnosed RA (cases per million people)				Incidence of additional algorithm diagnosed RA (cases per million people)				Percentage change in the doctor diagnosed incidence from the addition of algorithm diagnosed cases			
	18-44	45-64	65-74	75+	18-44	45-64	65-74	75+	18-44	45-64	65-74	75+
2000	63.3	597.1	1108.1	1081.7	23.8	168	103	48.5	37.6	28.1	9.3	4.5
2001	72.1	657.4	1253.1	1266.2	29.3	183.7	149.2	69.8	40.6	27.9	11.9	5.5
2002	79	699.2	1198.4	1203.6	33.5	197.6	141.2	94	42.5	28.3	11.8	7.8
2003	83.6	700.6	1493.8	1253.7	41.1	255	207.7	93.3	49.2	36.4	13.9	7.4
2004	102.6	896.9	1670.8	1381.3	37.5	242	211.5	118.3	36.6	27	12.7	8.6
2005	104.4	832.3	1590.2	1247.5	40.4	215.5	237	92.7	38.7	25.9	14.9	7.4
2006	113.9	763.8	1380.8	1118	36.7	227	191.7	87.4	32.2	29.7	13.9	7.8
2007	103.1	784.9	1375.2	962.8	29.7	191.4	228.7	111.4	28.8	24.4	16.6	11.6
2008	104.5	719.1	1315.5	916.3	36.5	214.8	209.9	92.3	34.9	29.9	16	10.1
2009	122.2	671	1224	851.9	33.8	201.2	184.6	82.5	27.7	30	15.1	9.7
2010	107.2	666.1	1032.2	717.9	27.3	182	198.7	64.8	25.5	27.3	19.2	9
2011	107.2	611.8	1160	709.1	26.3	156.2	221.2	80.3	24.5	25.5	19.1	11.3
2012	118.1	641.6	1164.4	735.7	30.9	149.9	161.5	81.7	26.2	23.4	13.9	11.1
2013	155.4	843.6	1583.6	1020.3	24.6	140.6	116.9	73.4	15.8	16.7	7.4	7.2
2014	155.4	897.8	1669.6	1065.7	19.7	112.3	129.2	63.7	12.7	12.5	7.7	6

4.5.7 Baseline comparison between doctor-diagnosed cases and algorithm-defined cases

We aimed to compare individuals baseline data extracted from Clinical Practice Research Datalink (CPRD) regarding different age, gender and serology test (rheumatoid factor) result groups. Number of participants and frequency are used to describe participants' characteristics. Odds ratios (ORs) and 95% confidence intervals (95%CI) are reported.

Of the 99683 adults (age≥18) in the CPRD dataset, 86893 of them are doctor-diagnosed RA cases and 12790 are additional algorithm identified RA cases (Table 46). In comparison, patients from 45-64 age groups account for the most of the cases both in doctor-diagnosed cases and algorithm-diagnosed groups (37.78% and 48.80% respectively). However, within algorithm diagnosed group, patients from 18-44 age group account for 25.94% of the total algorithm cases compared with 17.87% in the doctor-diagnosed group.

Among them, females (OR=1.20, 95% CI: 1.16- 1.25), increasing age, especially for 65-74 (OR=2.15, 95%CI=2.02- 2.28) and over 75 age group (OR=3.15, 95% CI: 2.94- 3.37) are significantly (P< 0.001) associated with doctor-diagnosed RA cases. In addition, nearly 55% RA patients (there are a large of number of patients' RF results missing) shown negative RF results. By contrast, all the additional RA cases identified by algorithm shown a high positive RF test results. Results are shown in Table 47.
In addition, we have found that within doctor-diagnosed RA cases, there are a number of patients identified by algorithm at an early age, we have further analysed this.

Table 46: Concordance between doctor- and algorithm diagnosis

Algorithm RA cases	Doctor diagnosis of RA		Total
	No	Yes	
No	0	86,893	86,893
Yes	12,790	0	12,790
Total	12,790	86,893	99,683

Table 47: Baseline values comparison and logistic regression analysis

Variable	Doctor diagnosis of RA	Additional algorithm diagnosis of RA	Logistic regression		
	Baseline	Baseline	Odds Ratio	95% CI	p-value
Age (agegrp)	Number (%)	Number (%)			
18-44	15390 (17.87%)	3318 (25.94%)	1.00		
45-64	32531 (37.78%)	6242 (48.80%)	1.12	[1.07- 1.17]	P<0.001
65-74	19392 (22.52%)	1944 (15.20%)	2.15	[2.02- 2.28]	P<0.001
Over 75	18800 (21.83%)	1286 (10.05%)	3.15	[2.94- 3.37]	P<0.001
Gender	Number (%)	Number (%)			
Male	25464 (29.31%)	4268 (33.37%)	1.00		
Female	61421 (70.69%)	8522 (66.63%)	1.20	[1.16- 1.25]	P<0.001
Serology	Number (%)	Number (%)			
Negative	5641 (54.97%)	0			
Low positive	720 (7.02%)	0			
High positive	3901 (38.01%)	12790 (100%)			

4.5.8 Doctor diagnosis delays

We have found that there are some RA cases identified by algorithm at an early age compared with doctor diagnosed age. The aim of this analysis is to determine potential doctor diagnosis delays.

Of the 99,683 individuals, there are 3,091 cases both identified by doctor and algorithm. The mean values of doctor diagnosed age and algorithm defined age are 60.19 and 57.68, respectively. There is statistically significant difference between these two age groups ($t = 20.07$, $p<0.001$). Of the 3,091 RA cases, nearly 63.99% (n=1,978) found doctor diagnosis delays. The histogram and Q-Q plot show potential delays by doctor diagnosis (Figure 7).

Early diagnosis and treatment of recent onset RA is a prime objective for clinical practitioners. In a multicentre European study, the median delay across the 10 centres from symptom onset to assessment was 24 weeks, with the percentage of patients seen within 12 weeks of symptom onset ranging from 8% to 42%.[34] The consequences of RA include severe and progressive joint damage as well as disability which lead to increased morbidity and mortality. The impacts of RA are linked with delays in diagnosis and control of inflammation and disease activity. Among a Dutch cohort of RA patients, 69% were assessed in ≥12 weeks; this was associated with a hazard ratio of 1.87 for not achieving DMARD-free remission and a 1.3 times higher rate of joint destruction over 6 years, as compared with assessment in <12 weeks.[35] In addition, all treatment options including monotherapy, combination DMARDs and biologics work better in early RA than in established conditions, which is explained by the concept of "window of opportunity".[36] relationships between symptom duration and favourable outcomes are not linear, and a point is reached after which the benefit gained by reducing time to treatment is lessened. Analysis of the primary outcome of DMARD-free sustained remission in two cohorts showed that the "window" appeared to start closing at 14.9 weeks.[37] Therefore, ability to define patients at the early course of the disease is significant in achieving remission and reduce the impacts.

The 2010 ACR criteria is designed to classify patients with early RA, which has shown an advantage in identifying early RA cases. However, a previous study indicates that specificity of these criteria is only 55% lower than the 1987 ACR criteria, and patients with systemic lupus erythematosus, OA and psoriatic arthritis may be classified as RA.[34] This would be seen as a potential limitation of current study.

Thus, it is worth to compare the results with patients on DMARDs without other indications except RA in order to further validate the algorithm diagnosis method.

Table 48: Comparison of diagnosed age regarding doctor diagnosis and algorithm defined diagnosis

	Observations	Mean	Standard Deviation	95% CI	P value
Doctor diagnosed age	3091	60.19	14.69	[59.67, 60.71]	P<0.001
Algorithm defined age	3091	57.68	14.82	[57.16, 58.20]	P<0.001

Table 49: Diagnosis delays between doctor diagnosed cases and algorithm diagnosed cases

	Frequency	Percentage
Algorithm delay	527	17.05%
No delay	586	18.96%
Doctor diagnosis delay	1978	63.99%
Total	3091	100.00

Figure 6: Distribution of doctor diagnosis delays

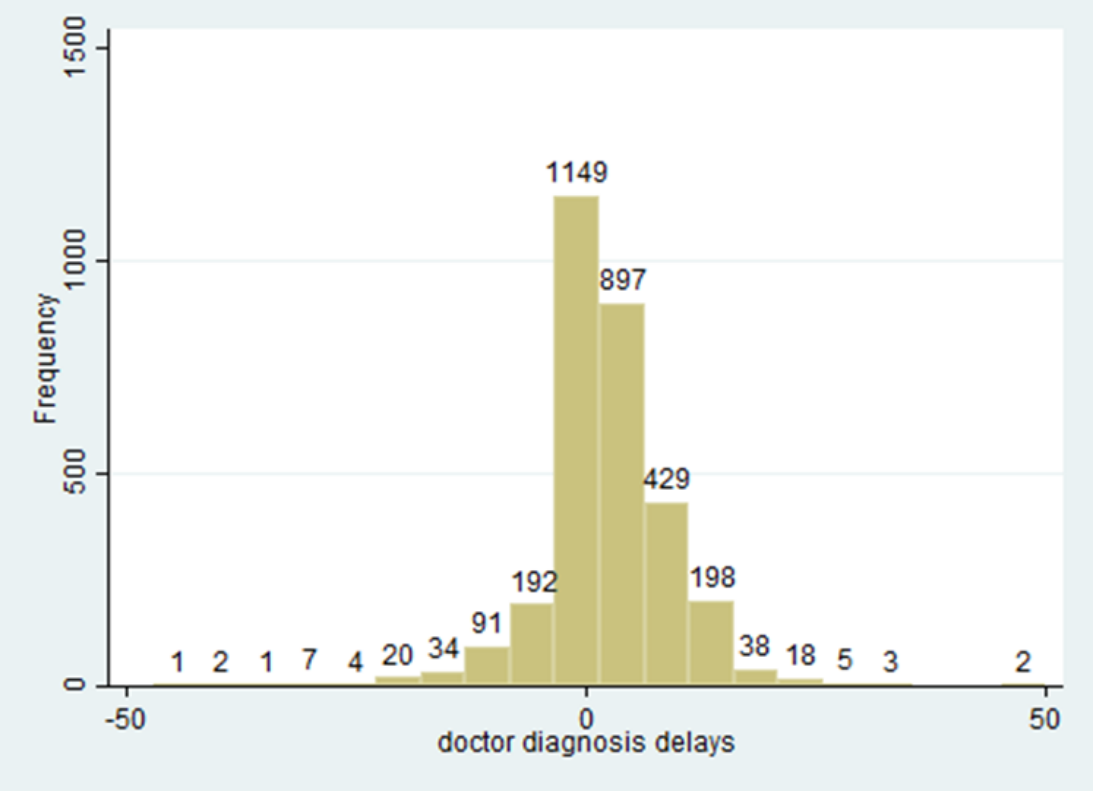
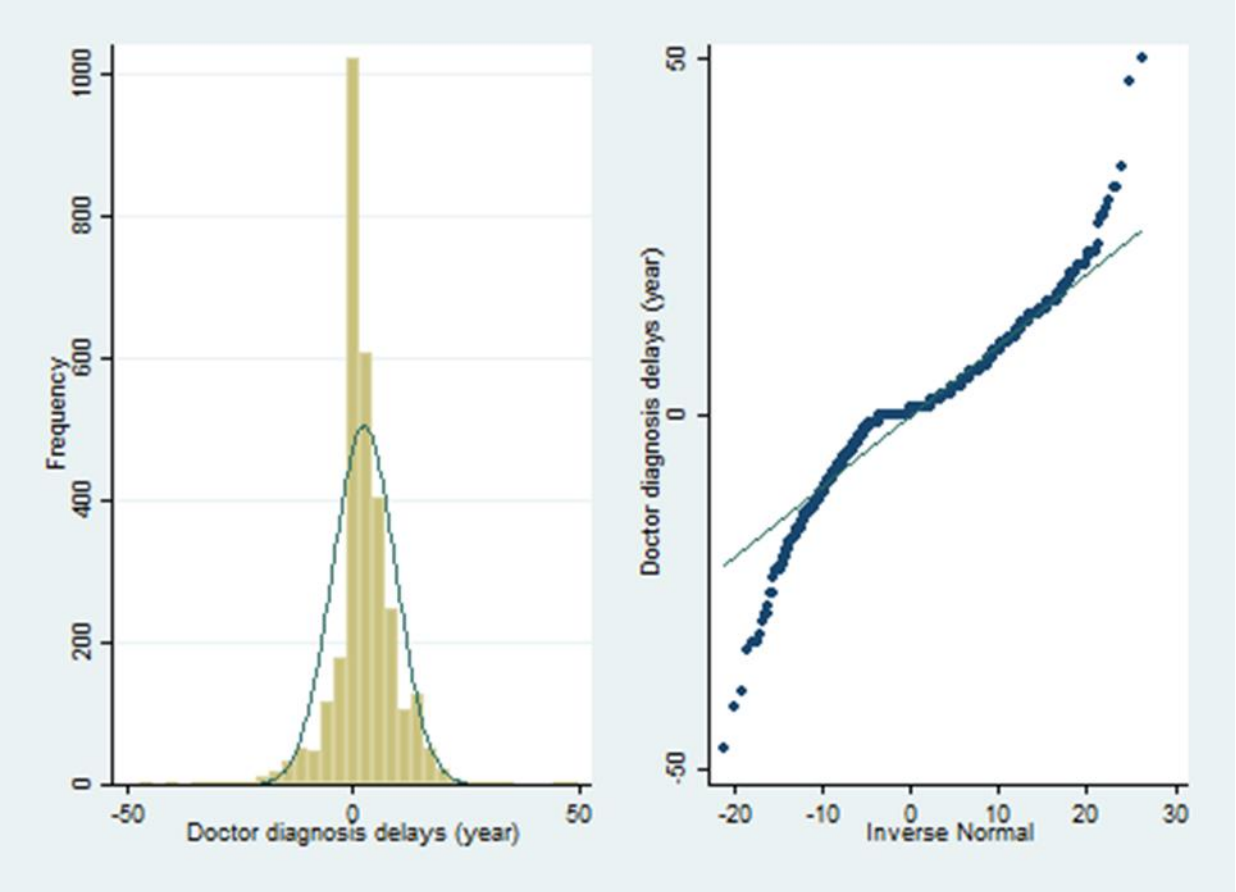


Figure 7: Histogram and Q-Q plot



4.6 Regression modelling using CPRD data

4.6.1 Missing data

CPRD data source may not include patient's data in terms of all the demographic aspects, such as ethnicity, smoking, alcohol consumption and BMI. There is some missing data in the above areas (Table 50), and different methods were used to deal with missing data. For ethnicity, missing data were considered as "White population". Multiple imputation was used to replace missing values for BMI, smoking status, alcohol consumption and deprivation. Table 51 shows the baseline characteristics of patients (both identified RA cases and non-RA cases) that included in the model. The characteristics of these five groups are relatively similar, despite that there is a greater number of younger populations in the controls group. Male patients were less than females for RA patients compared with non-RA cases.

Table 50: proportions of missing data

Predictor variables	Total
Total number of respondents	455,898
Gender	
Male	238,407 (44.03%)
Female	255,148 (55.97%)
Missing	0%
Age group	
18-44	227,874 (49.98%)
45-64	121,320 (26.61%)
65-74	48,289 (10.59%)
>75	58,415 (12.81%)
Missing	0%
Alcohol	
Non-drinker	47,026 (10.32%)
Light (<15 units per week)	186,443 (40.90%)
Moderate (14-42 units per week)	35,249 (7.73%)
Heavy (>42 units per week)	8,635 (1.89%)
Missing	178,545 (39.16%)
Ethnicity	
White	69,700 (15.29%)
Mixed	2,486 (0.55%)
Asian	6,781 (1.49%)
Black	4,092 (0.90%)
Other	2,976 (0.65%)
Missing	369,863 (81.13%)
BMI	
Underweight (<18.5)	11,645 (2.55%)
Normal (18.5-25)	138,274 (30.33%)
Overweight (25-30)	105,761 (23.20%)
Obese (>30)	72,049 (15.80%)
Missing	128,169 (28.11%)
Smoking	

Predictor variables	Total
Non-smoker	198,552 (43.55%)
Ex-smoker	84,474 (18.53%)
Smoker	83,562 (18.33%)
%Missing	89,310 (19.59%)
Deprivation	
1 (least deprived)	17,739 (3.89%)
2	21,195 (4.65%)
3	23,633 (5.18%)
4	25,785 (5.66%)
5 (most deprived)	22,316 (4.89%)
Missing	345,230 (75.73%)

4.6.2 Baseline descriptive characteristics of CPRD patients

Table 51 shows the baseline characteristics of patients (both identified RA cases and non-RA cases) that included in the model. The characteristics of these five groups are relatively similar, despite that there is a greater number of younger populations in the controls group. Male patients were less than females for RA patients compared with non-RA cases.

Table 51: Baseline characteristics of patients involved in the logistic regression model

Predictor variables	Doctor diagnosed cases	HES cases	Algorithm defined cases	DMARDs cases	Controls	Total
Total number of respondents	82,736	791	12,762	5,303	354,306	455,898
Gender						
Male	24,577 (29.71%)	242 (30.59%)	4,251 (33.31%)	1,769 (33.36%)	169,911 (47.96%)	200,750 (44.03%)
Female	58,159 (70.29%)	549(69.41%)	8,511 (66.69%)	3,534 (66.64%)	184,395 (52.04%)	255,148 (55.97%)
Total	82,736	791	12,762	5,303	354,306	455,898
Age group						
18-44	15,003 (18.13%)	126 (15.93%)	3,314 (25.97%)	1,432 (27.00%)	207,999 (58.71%)	227,874 (49.98%)
45-64	31,610 (38.21%)	247 (31.23%)	6,220 (48.74%)	2,135 (40.26%)	81,108 (22.89%)	121,320 (26.61%)
65-74	18,865 (22.80%)	159 (20.10%)	1,948 (15.26%)	921 (17.37%)	26,396 (7.45%)	48,289 (10.59%)
>75	17,258 (20.86%)	259 (32.74%)	1,280 (10.03%)	815(15.37%)	38,803 (10.95%)	58,415 (12.81%)
Total	82,736	791	12,762	5,303	354,306	455,898
Alcohol						
Non-drinker	21,931(36.76%)	66 (16.54%)	1,317 (12.02%)	639 (15.71%)	34,600 (16.34%)	47,026 (16.96%)
Light (<15 units per week)	2,808 (4.71%)	288 (72.18%)	7,918 (72.28%)	2,953 (72.59%)	140,661 (66.44%)	186,443 (67.22%)
Moderate (14-42 units per week)	19,452 (32.60%)	35 (8.77%)	1,391(12.70%)	407(10.00%)	29,205 (13.79%)	35,249 (12.71%)
Heavy (>42 units per week)	15,469 (25.93%)	10 (2.51%)	329 (3.00%)	69(1.70%)	7,249 (3.42%)	8,635 (3.11%)

Predictor variables	Doctor diagnosed cases	HES cases	Algorithm defined cases	DMARDs cases	Controls	Total
Total	50,216	399	10,955	4,068	211,715	277,353
Ethnicity						
White	80,379 (97.15%)	769 (97.22%)	12,083 (94.68%)	5,046 (95.15%)	341,286 (96.33%)	439,563 (96.42%)
Mixed	530 (0.64%)	7 (0.88%)	153 (1.20%)	59 (1.11%)	1,737 (0.49%)	2,486 (0.55%)
Asian	1,049 (1.27%)	9 (1.14%)	319 (2.50%)	97 (1.83%)	5,307 (1.50%)	6,781 (1.49%)
Black	419 (0.51%)	5 (0.63%)	137 (1.07%)	62 (1.17%)	3,469 (0.98%)	4,092 (0.90%)
Other	359 (0.43%)	1 (0.13%)	70 (0.55%)	39 (0.74%)	2,507 (0.71%)	2,976 (0.65%)
Total	82,736	791	12,762	5,303	354,306	455,898
BMI						
Underweight (<18.5)	2,808 (4.71%)	21 (4.27%)	209 (1.70%)	148 (3.02%)	8,459 (3.38%)	11,645 (3.55%)
Normal (18.5-25)	21,931 (36.76%)	154 (31.30%)	3,563 (29.04%)	1587 (32.37%)	111,039 (44.34%)	138,274 (42.19%)
Overweight (25-30)	19,452 (32.60%)	155 (31.50%)	4,307 (35.10%)	1,611 (32.86%)	80,236 (32.04%)	105,761 (32.27%)
Obese (>30)	15,469 (25.93%)	162 (32.93%)	4,190 (34.15%)	1,556 (31.74%)	50,672 (20.24%)	72,049 (21.98%)
Total	59,660	492	12,269	4,902	250,406	327,729
Smoking						
Non-smoker	32,020 (49.81%)	300 (57.36%)	6,771 (53.21%)	2,840 (54.24%)	156,621 (55.18%)	198,552 (54.16%)
Ex-smoker	20,958 (32.60%)	145 (27.72%)	3,835 (30.14%)	1,577 (30.12%)	57,959 (20.42%)	84,474 (23.04%)
Smoker	11,307 (17.59%)	78 (14.91%)	2,120 (16.66%)	819 (15.64%)	69,238 (24.40%)	83,562 (22.79%)
Total	64,285	523	12,726	5,236	283,818	366,588
Deprivation						
1 (least deprived)	12,964 (15.67%)	78 (13.04%)	2,032 (16.24%)	47 (14.60%)	2,618 (18.06%)	17,739 (16.03%)
2	15,677 (18.95%)	147 (24.58%)	2,375 (18.98%)	72 (22.36%)	2,924 (20.17%)	21,195 (19.15%)
3	17,724 (21.42%)	124 (20.74%)	2,551 (20.39%)	71 (22.05%)	3,163 (21.81%)	23,633 (21.35%)
4	19,543 (23.62%)	123 (20.57%)	2,860 (22.86%)	66 (20.50%)	3,193 (22.02%)	25,785 (23.30%)
5 (most deprived)	16,828 (20.34%)	126 (21.07%)	2,694 (21.53%)	66 (20.50%)	2,602 (17.94%)	22,316 (20.16%)
Total	82,736	598	12,512	322	14,500	110,668

4.6.3 CPRD univariate logistic analysis

Table 52 shows the results of univariate logistic models for individual risk factors and the outcome.

Table 52: Univariate logistic model 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			
18-44	1.00		
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]
Alcohol			
Non-drinker	1.00		
Light (<15 units per week)	0.906	<0.001	[0.886 – 0.927]
Moderate (14-42 units per week)	0.576	<0.001	[0.557 - 0.596]
Heavy (>42 units per week)	0.532	<0.001	[0.501 - 0.566]
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]

4.6.4 Multivariate logistic analysis

We went through an extensive model fitting process to compare the performance of different models that included RA patients identified by different methods. Table 53- Table 60 below show the details of multivariate model fitting, and shows model M6 which is the model we finally selected.

Table 53: M1- Logistic regression model including patients only with CPRD doctor-diagnosed RA

Predictor variables	Odds Ratio	P>t	[95% Conf. Interval]
Gender			
Male	1.00		
Female	2.208	<0.001	[2.165 - 2.251]
Age group			
18-44	1.00		
45-64	5.137	<0.001	[5.025 - 5.251]
65-74	9.249	<0.001	[9.010 - 9.494]
>75	5.681	<0.001	[5.537- 5.828]
Alcohol			
Non-drinker	1.00		
Light (<15 units per week)	0.985	0.305	[0.955 - 1.016]
Moderate (14-42 units per week)	0.801	<0.001	[0.759 - 0.844]
Heavy (>42 units per week)	0.750	<0.001	[0.690 - 0.815]
Ethnicity			
White	1.00		
Mixed	1.423	<0.001	[1.278 - 1.586]
Asian	1.437	<0.001	[1.336 - 1.546]
Black	0.813	<0.001	[0.729 - 0.906]
Other	0.998	0.976	[0.886 - 1.124]
BMI			
Underweight (<18.5)	1.262	<0.001	[1.208 - 1.318]
Normal (18.5-25)	1.00		
Overweight (25-30)	1.117	<0.001	[1.091 - 1.143]
Obese (>30)	1.343	<0.001	[1.307 - 1.381]
Smoking			
Non-smoker	1.00		
Ex-smoker	1.583	<0.001	[1.548 - 1.620]
Smoker	1.144	<0.001	[1.118 - 1.172]
Deprivation			
1 (least deprived)	1.00		
2	1.123	<0.001	[1.085 - 1.162]
3	1.130	0.032	[1.016 - 1.257]
4	1.272	0.003	[1.133 - 1.428]
5 (most deprived)	1.374	<0.001	[1.291 - 1.462]
_cons	0.312	<0.001	[0.029 - 0.034]

Table 54: M2- Logistic regression model including patients with CPRD doctor diagnosed RA and HES RA diagnosis

Predictor variables	Odds Ratio	P>t	[95% Conf. Interval]
Gender			
Male	1.00		
Female	2.209	<0.001	[2.168 – 2.250]
Age group			
18-44	1.00		
45-64	5.128	<0.001	[5.017 – 5.243]
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]

Table 55: M3- Logistic regression model including patients with CPRD doctor diagnosed RA, HES RA diagnosis and algorithm defined RA cases

Predictor variables	Odds Ratio	P>t	[95% Conf. Interval]
Gender			
Male	1.00		
Female	2.206	<0.001	[2.169 - 2.245]
Age group			
18-44	1.00		
45-64	4.989	<0.001	[4.889 – 5.090]
65-74	8.431	<0.001	[8.227 - 8.641]
>75	5.129	<0.001	[5.005 – 5.256]
Alcohol			
Non-drinker	1.00		
Light (<15 units per week)	1.061	0.001	[1.027 – 1.096]
Moderate (14-42 units per week)	0.928	<0.001	[0.898 - 0.959]
Heavy (>42 units per week)	0.897	0.014	[0.825 - 0.975]
Ethnicity			
White	1.00		
Mixed	1.555	<0.001	[1.409 - 1.716]
Asian	1.594	<0.001	[1.492 - 1.703]
Black	0.894	0.024	[0.811 - 0.985]
Other	1.009	0.868	[0.904 - 1.127]
BMI			
Underweight (<18.5)	1.204	<0.001	[1.145 - 1.265]
Normal (18.5-25)	1.00		
Overweight (25-30)	1.175	<0.001	[1.149 - 1.201]
Obese (>30)	1.477	<0.001	[1.442 – 1.513]
Smoking			
Non-smoker	1.00		
Ex-smoker	1.539	<0.001	[1.508 -1.571]
Smoker	1.098	<0.001	[1.070 -1.128]
Deprivation			
1 (least deprived)	1.00		
2	1.114	0.015	[1.031 – 1.205]
3	1.139	0.002	[1.070 – 1.212
4	1.246	0.001	[1.148 – 1.352]
5 (most deprived)	1.373	<0.001	[1.308 – 1.442]
_cons	0.034749	<0.001	[0.032 – 0.037]

Table 56: M4- Logistic regression model including patients with CPRD doctor diagnosed RA, HES RA diagnosis and combination DMARDs cases

Predictor variables	Odds Ratio	P>t	[95% Conf. Interval]
Gender			
Male	1.00		
Female	2.208	<0.001	[2.168 - 2.250]
Age group			
18-44	1.00		
45-64	5.128	<0.001	[5.017 - 5.243]
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		
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)	1.00		
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]

Table 57: M5- Logistic regression model including patients with CPRD doctor diagnosed RA, HES RA diagnosis and DMARDs cases

Predictor variables	Odds Ratio	P>t	[95% Conf. Interval]
Gender			
Male	1.00		
Female	2.209	<0.001	[2.168-2.250]
Age group			
18-44	1.00		
45-64	5.128	<0.001	[5.017-5.243]
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	[0.731-0.908]
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		
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)	1.00		
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]

Table 58: M6- Logistic regression model including patients with CPRD doctor diagnosed RA, HES RA diagnosis, algorithm identified RA cases and DMARDs cases

Predictor variables	Odds Ratio	P>t	[95% Conf. Interval]
Gender			
Male	1.00		
Female	2.207	<0.001	[2.169-2.246]
Age group			
18-44	1.00		
45-64	4.999	<0.001	[4.897-5.103]
65-74	8.471	<0.001	[8.260-8.687]
>75	5.146	<0.001	[5.020-5.275]
Alcohol			
Non-drinker	1.00		
Light (<15 units per week)	1.065	<0.001	[1.038-1.093]
Moderate (14-42 units per week)	0.934	0.003	[0.895-0.976]
Heavy (>42 units per week)	0.894	<0.001	[0.846-0.944]
Ethnicity			
White	1.00		
Mixed	1.561	<0.001	[1.413-1.723]
Asian	1.599	<0.001	[1.496-1.709]
Black	0.896	0.027	[0.813-0.988]
Other	1.010	0.856	[0.904-1.129]
BMI			
Underweight (<18.5)	1.194	<0.001	[1.142-1.248]
Normal (18.5-25)	1.00		
Overweight (25-30)	1.175	<0.001	[1.149-1.202]
Obese (>30)	1.476	<0.001	[1.443-1.509]
Smoking			
Non-smoker	1.00		
Ex-smoker	1.526	<0.001	[1.483-1.571]
Smoker	1.095	<0.001	[1.062-1.128]
Deprivation			
1 (least deprived)	1.00		
2	1.117	0.032	[1.014-1.230]
3	1.124	0.017	[1.031-1.227]
4	1.244	0.002	[1.130-1.370]
5 (most deprived)	1.372	<0.001	[1.261-1.494]
_cons	0.035	<0.001	[0.032-0.037]

Table 59: M7- Logistic regression model including patients only with algorithm defined RA cases

Predictor variables	Odds Ratio	P>t	[95% Conf. Interval]
Gender			
Male	1.00		
Female	2.105779	<0.001	[2.023-2.192]
Age group			
18-44	1.00		
45-64	4.384	<0.001	[4.195-4.582]
65-74	4.457	<0.001	[4.202-4.728]
>75	2.195	<0.001	[2.052-2.347]
Alcohol			
Non-drinker	1.00		
Light (<15 units per week)	1.660	<0.001	[1.555-1.772]
Moderate (14-42 units per week)	1.836	<0.001	[1.683-2.002]
Heavy (>42 units per week)	1.833	<0.001	[1.608-2.089]
Ethnicity			
White	1.00		
Mixed	2.408	<0.001	[2.024-2.866]
Asian	2.609	<0.001	[2.314-2.941]
Black	1.388	<0.001	[1.163-1.658]
Other	1.155	0.247	[0.905-1.475]
BMI			
Underweight (<18.5)	0.714	<0.001	[0.618-0.824]
Normal (18.5-25)	1.00		
Overweight (25-30)	1.488	<0.001	[1.421-1.559]
Obese (>30)	2.128	<0.001	[2.029-2.232]
Smoking			
Non-smoker	1.00		
Ex-smoker	1.297	<0.001	[1.243-1.354]
Smoker	0.896	<0.001	[0.850-0.944]
Deprivation			
1 (least deprived)	1.00		
2	1.033	0.62	[0.891-1.198]
3	1.017	0.768	[0.893-1.160]
4	1.124	0.069	[0.988-1.279]
5 (most deprived)	1.292	0.004	[1.119-1.493]
_cons	0.0041	<0.001	[0.004-0.005]

Table 60: M8- Logistic regression model including patients with algorithm identified RA cases and DMARDs cases

Predictor variables	Odds Ratio	P>t	[95% Conf. Interval]
Gender			
Male	1.00		
Female	2.114	<0.001	[2.028-2.203]
Age group			
18-44	1.00		
45-64	4.396	<0.001	[4.204-4.598]
65-74	4.500	<0.001	[4.239-4.776]
>75	2.212	<0.001	[2.066-2.369]
Alcohol			
Non-drinker	1.00		
Light (<15 units per week)	1.667	<0.001	[1.575-1.766]
Moderate (14-42 units per week)	1.860	<0.001	[1.705-2.029]
Heavy (>42 units per week)	1.825	<0.001	[1.579-2.110]
Ethnicity			
White	1.00		
Mixed	2.443	<0.001	[2.053-2.908]
Asian	2.660	<0.001	[2.358-3.001]
Black	1.400	<0.001	[1.171-1.673]
Other	1.156	0.251	[0.903-1.479]
BMI			
Underweight (<18.5)	0.713	<0.001	[0.616-0.824]
Normal (18.5-25)	1.00		
Overweight (25-30)	1.488	<0.001	[1.420-1.558]
Obese (>30)	2.114	<0.001	[2.012-2.220]
Smoking			
Non-smoker	1.00		
Ex-smoker	1.297	<0.001	[1.242-1.353]
Smoker	0.904	<0.001	[0.857-0.954]
Deprivation			
1 (least deprived)	1.00		
2	1.046	0.317	[0.953-1.149]
3	1.003	0.931	[0.938-1.073]
4	1.137	0.006	[1.043-1.240]
5 (most deprived)	1.321	0.005	[1.126-1.551]
_cons	0.004	<0.001	[0.004-0.004]

4.6.5 ROC curves

We next examined the receiver operating characteristics (ROC) curves for the various models. The best ROC curve which predicts data perfectly will touch the top-left corner of the plot (area 1.0), and the larger the area under the ROC curve the better the prediction. An area of 0.5 signifies a prediction no better than chance. The results are summarised in Table 61.

Table 61: receiver operating characteristics (ROC) curves for the various CPRD models

Model description	Model	ROC area	SE	95% CI
Logistic regression model including patients only with CPRD doctor diagnosed RA	M1	0.7661	0.0004	[0.765 - 0.767]
Logistic regression model including patients with CPRD doctor diagnosed RA and HES RA diagnosis	M2	0.7661	0.0004	[0.765 - 0.767]
Logistic regression model including patients with CPRD doctor diagnosed RA, HES RA diagnosis and algorithm defined RA cases	M3	0.7600	0.0004	[0.759 - 0.761]
Logistic regression model including patients with CPRD doctor diagnosed RA, HES RA diagnosis and combination DMARDs cases	M4	0.7661	0.0004	[0.765 - 0.767]
Logistic regression model including patients with CPRD doctor diagnosed RA, HES RA diagnosis and DMARDs cases	M5	0.7661	0.0004	[0.765 - 0.767]
Logistic regression model including patients with CPRD doctor diagnosed RA, HES RA diagnosis, algorithm identified RA cases and DMARDs cases	M6	0.7599	0.0004	[0.759 - 0.761]
Logistic regression model including patients only with algorithm defined RA cases	M7	0.7452	0.0009	[0.742 – 0.747]
Logistic regression model including patients with algorithm identified RA cases and DMARDs cases	M8	0.7456	0.0009	[0.744 – 0.747]

Figure 8: ROC curves-M1

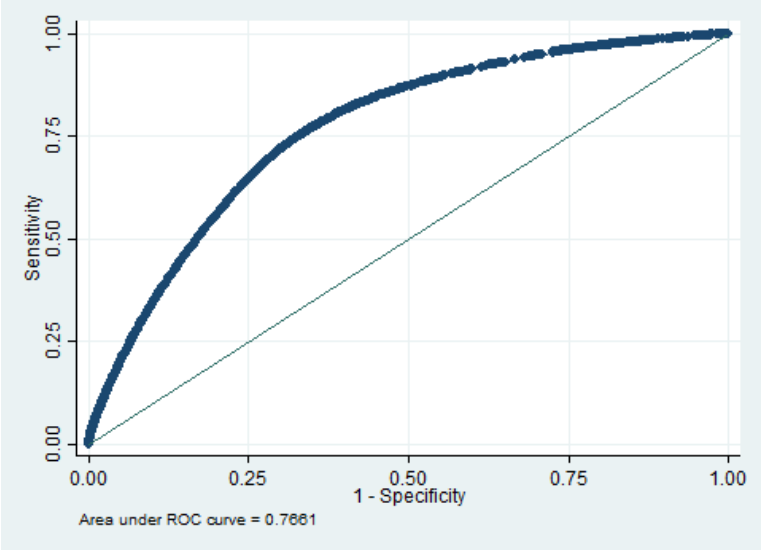


Figure 9: ROC curves-M2

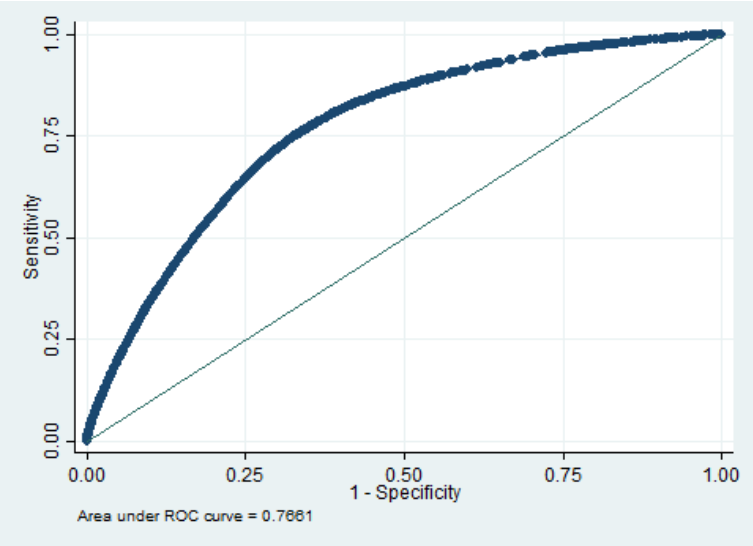


Figure 10: ROC curves-M3

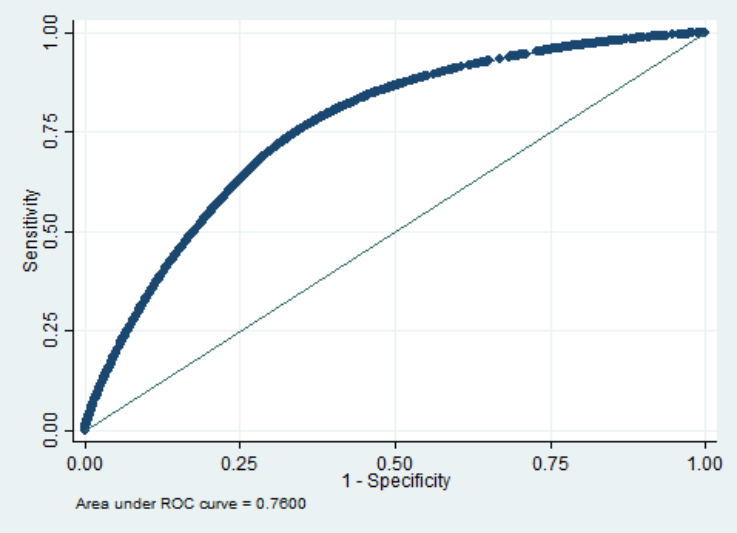


Figure 11: ROC curves-M4

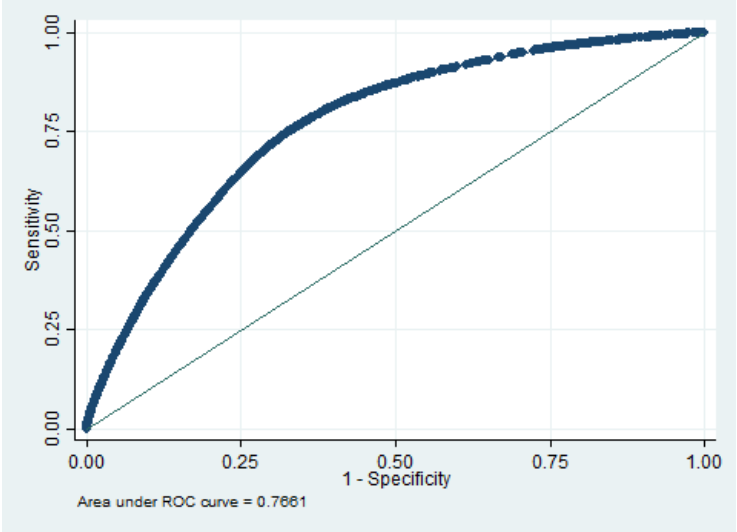


Figure 12: ROC curves-M5

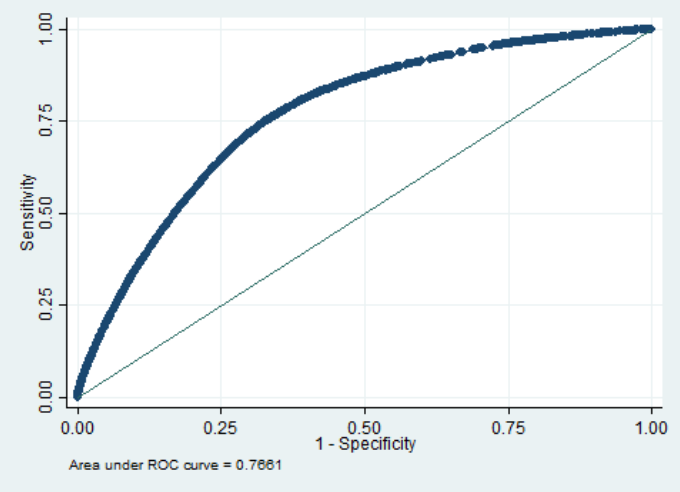


Figure 13: ROC curves-M6

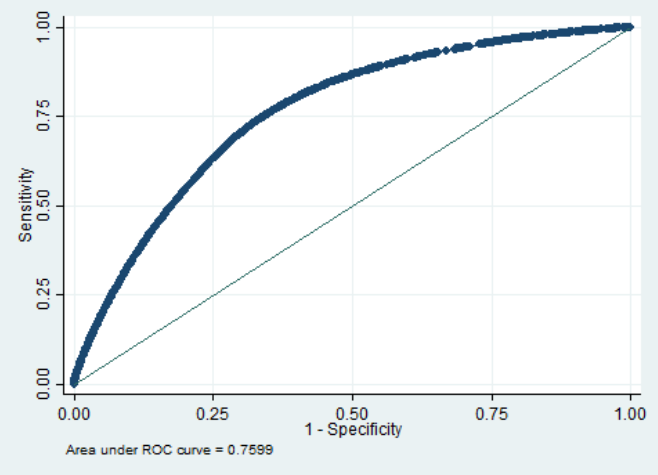


Figure 14: ROC curves-M7

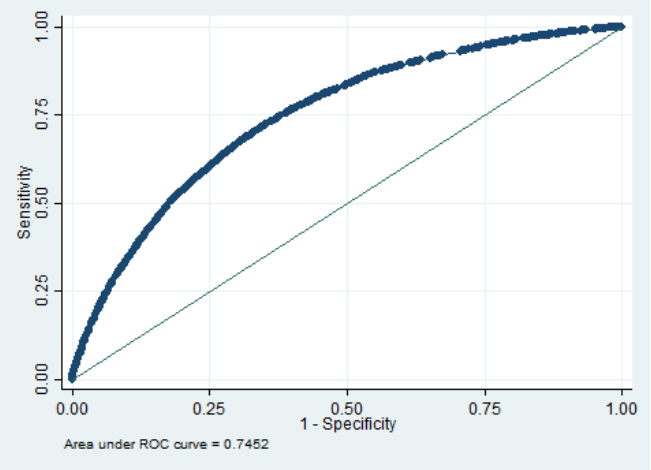
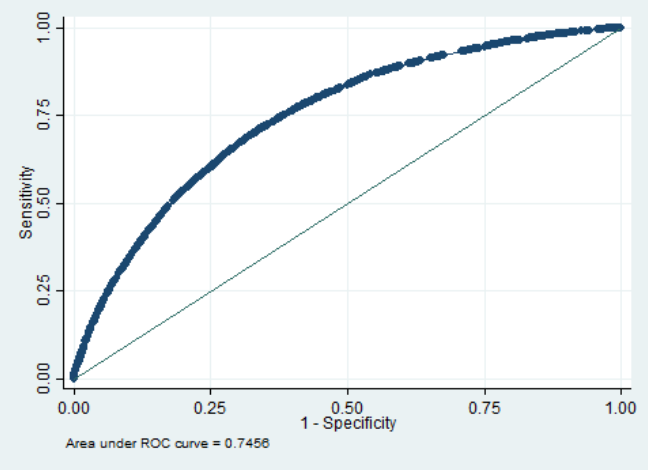


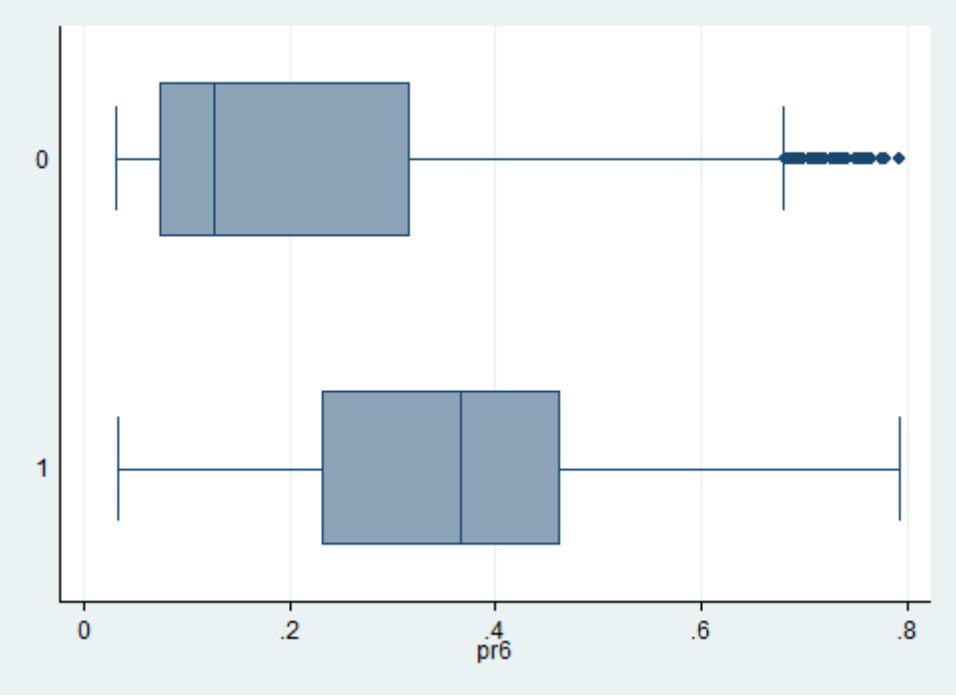
Figure 15: ROC curves-M8



4.6.6 Probability and sensitivity/specificity analysis

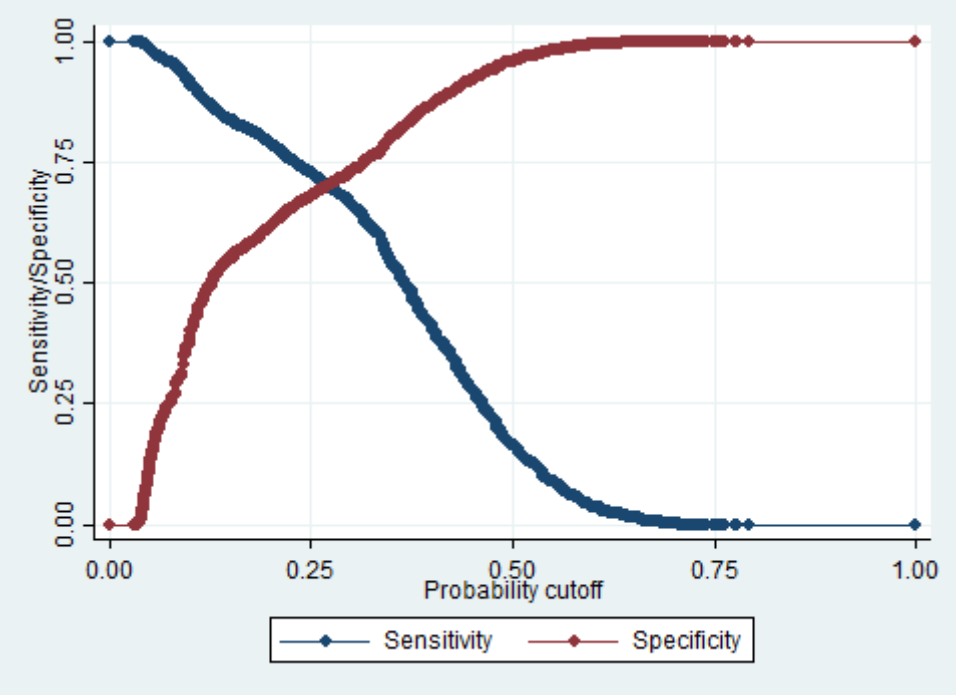
We could use the automatic stepwise forward model to predict the probability of individual being RA case in CPRD data set. In Figure 16 the two box plots show the predicted probability of people with RA caseness among the non-RA and RA groups. Since we have a binary response model, we can choose a cut-off point on the predicted probability to separate the predicted RA cases (with higher predicted probability) from the predicted non-RA 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-RA people being classified as predicted severe RA cases, or RA cases being classified as predicted non-RA cases. Therefore, we use sensitivity and specificity plots to help with this decision.

Figure 16: Predicted probabilities of being RA case



The sensitivity/specificity versus probability cut-off plot shows us the corresponding sensitivity and specificity in each possible probability cut-off point (See Figure 17). Higher sensitivity would usually yield low specificity and vice versa, the rule of thumb is to choose a cut-off probability to maximize both. We choose the cut-off probability where sensitivity and specificity lines cross.

Figure 17: Sensitivity/specificity versus probability cut-off



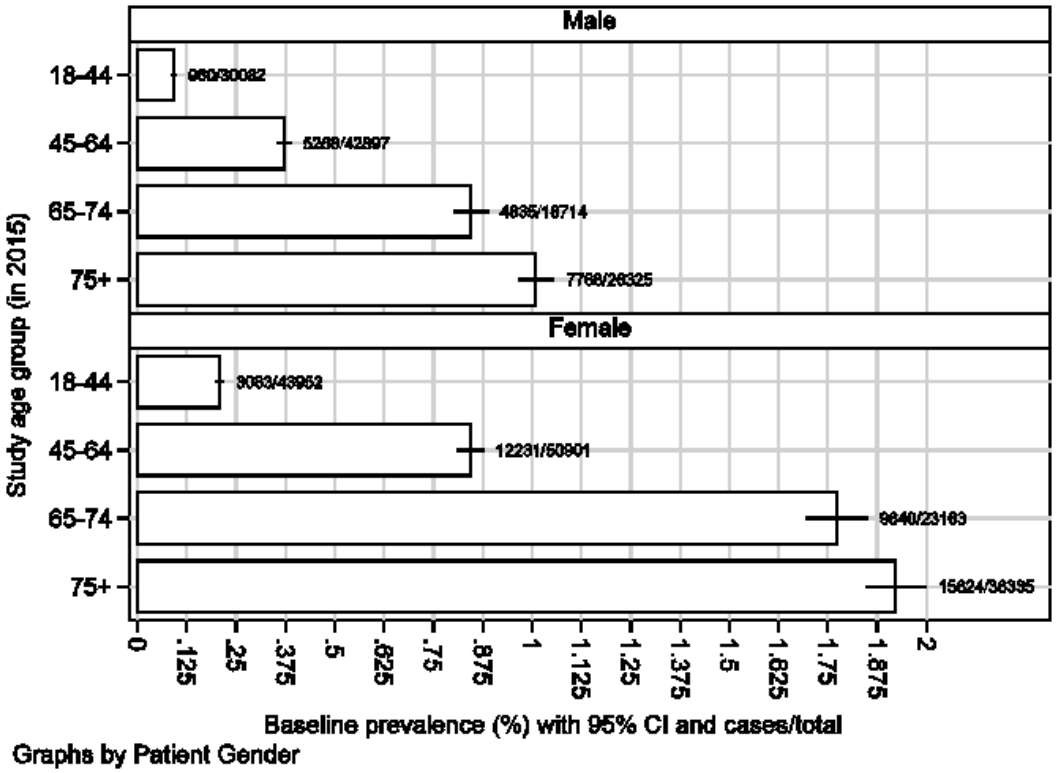
4.7 Population RA prevalence using local estimation Method 2, sampling-probability weights

The bootstrap local estimation Method 1 does not generate national/whole population prevalence estimates. However because it is a two-step process Method 2, sampling-probability weights, does do so. These results before probability weighting are shown in Table 62 and Figure 18.

Table 62: Baseline prevalence of RA (%) by gender and age

Age group	Cases/total	Prevalence (%)	(95%	CI)
Male:				
18-44	960/30082	0.092	(0.085,	0.099)
45-64	5268/42897	0.372	(0.355,	0.390)
65-74	4835/18714	0.846	(0.804,	0.889)
75+	7768/26325	1.010	(0.965,	1.056)
Female:				
18-44	3083/43952	0.209	(0.199,	0.220)
45-64	12231/50901	0.844	(0.811,	0.878)
65-74	9640/23163	1.772	(1.697,	1.851)
75+	15624/36335	1.922	(1.848,	1.999)

Figure 18: Baseline prevalence of RA (%) by gender and age

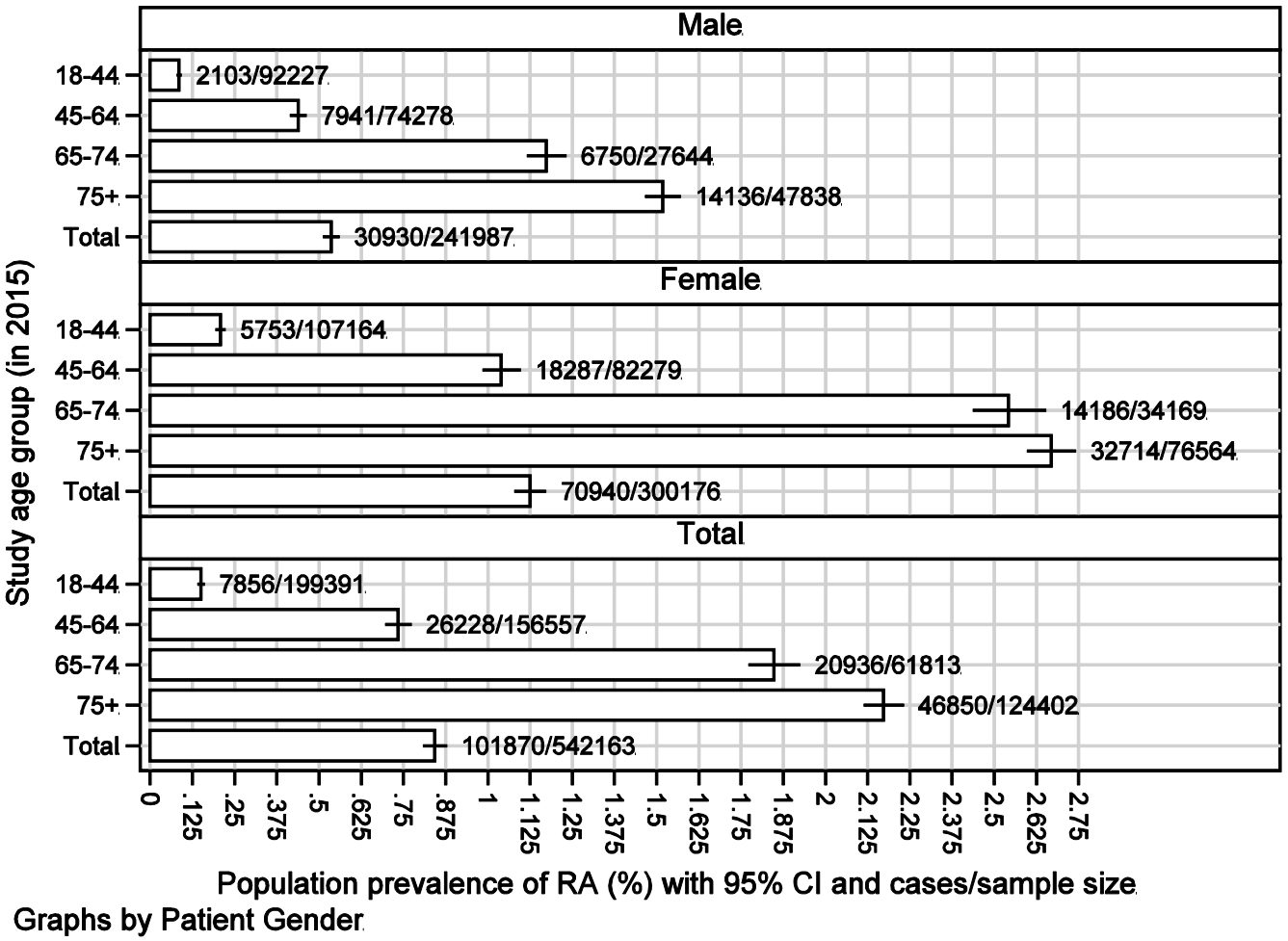


These results <u>after</u> probability weighting are shown in Table 63 and Figure 19 below.

Table 63: Population RA prevalence (%) by gender and age group

Age group	Cases/sample size	Prevalence (%)	(95%	CI)
Male				
18-44	2103/92227	0.086	(0.079,	0.093)
45-64	7941/74278	0.438	(0.414,	0.463)
65-74	6750/27644	1.173	(1.117,	1.233)
75+	14136/47838	1.518	(1.466,	1.572)
Total	30930/241987	0.536	(0.512,	0.560)
Female				
18-44	5753/107164	0.208	(0.194,	0.223)
45-64	18287/82279	1.039	(0.984,	1.097)
65-74	14186/34169	2.542	(2.437,	2.652)
75+	32714/76564	2.668	(2.597,	2.742)
Total	70940/300176	1.124	(1.079,	1.172)
Total				
18-44	7856/199391	0.151	(0.141,	0.161)
45-64	26228/156557	0.734	(0.697,	0.773)
65-74	20936/61813	1.847	(1.773,	1.925)
75+	46850/124402	2.172	(2.114,	2.232)
Total	101870/542163	0.843	(0.809,	0.878)

Figure 19: Population RA prevalence (%) by gender and age group



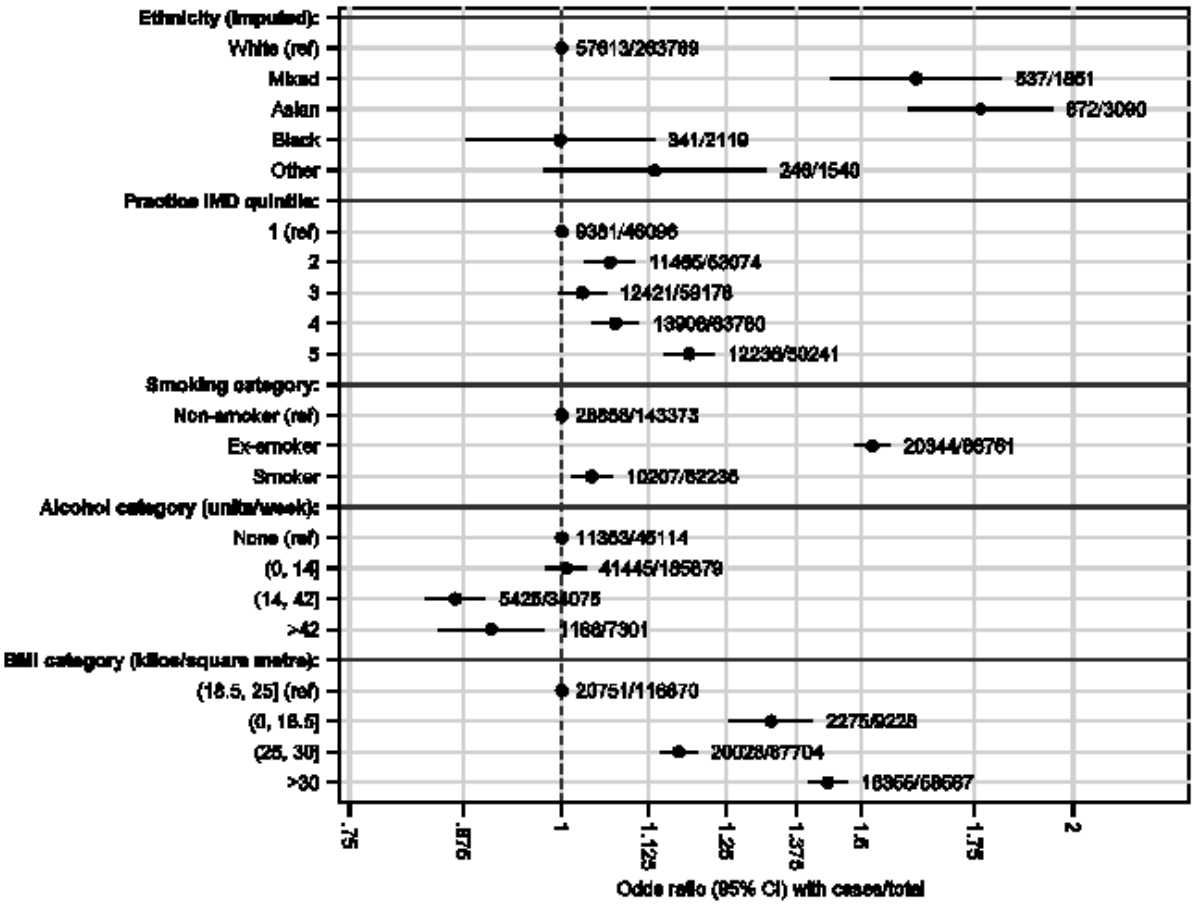
Finally, we fitted a logistic regression model of RA case status with respect to predictors using the probability weights method. These ORs are shown in Table 64 and Figure 20.

Table 64: Odds ratios for RA with respect to non-reference levels of risk factors

Age group	Cases/total	OR	(95%	CI)	P
Ethnicity (imputed):					
White	57613/263769	1.000	(ref)		
Mixed	537/1851	1.617	(1.441,	1.815)	3.4×10^{-16}
Asian	672/3090	1.764	(1.602,	1.944)	1.2×10^{-30}
Black	341/2119	0.997	(0.878,	1.133)	0.97
Other	246/1540	1.134	(0.976,	1.318)	0.1
Practice IMD quintile:					
1	9381/46096	1.000	(ref)		
2	11465/53074	1.067	(1.032,	1.104)	0.00014
3	12421/59178	1.028	(0.995,	1.062)	0.1
4	13906/63780	1.075	(1.041,	1.111)	0.000011
5	12236/50241	1.189	(1.149,	1.229)	1.1×10^{-23}
Smoking category:					

Age group	Cases/total	OR	(95%	CI)	P
Non-smoker	28858/143373	1.000	(ref)		
Ex-smoker	20344/66761	1.524	(1.488,	1.560)	$1.6x10^{-268}$
Smoker	10207/62235	1.042	(1.013,	1.071)	0.004
Alcohol category (units/week):					
None	11353/45114	1.000	(ref)		
(0, 14]	41445/185879	1.006	(0.980,	1.033)	0.65
(14, 42]	5425/34075	0.866	(0.832,	0.901)	$2.3x10^{-12}$
>42	1186/7301	0.909	(0.846,	0.977)	0.0095
BMI category (kilos/square metre):					
(18.5, 25]	20751/116870	1.000	(ref)		
(0, 18.5]	2275/9228	1.328	(1.256,	1.404)	$2.1x10^{-23}$
(25, 30]	20028/87704	1.172	(1.144,	1.201)	$7.7x10^{-38}$
>30	16355/58567	1.434	(1.397,	1.472)	$2.1x10^{-159}$

Figure 20: Odds ratios for RA with respect to non-reference levels of risk factors



4.8 Validation of local estimates

4.8.1 Internal validation of local estimates

In the CPRD dataset we used for the local estimates, we identified a total of 101,870 RA registered and possible cases. After dropping cases with a death date (N=23,904), there were 77,966 cases. The detailed results of crude regional prevalence is shown in Table 65. The average prevalence in the aggregated estimates is higher than that in the derivation dataset. Since the estimates are based on the prevalence of risk factors in each practice, this could occur because CPRD practices differ systematically from the other practices in each Region in terms of risk factors in their populations. The estimates are on average about 27% higher than those in the CPRD practices i.e. an average prevalence of 0.84% vs 0.56%.

4.8.2 External validation of local estimates

Table 66 compares practice-level and aggregate numbers and prevalence derived from the local practice-level estimates with corresponding QOF register data for England Regions. The bottom row shows the percentage difference between the local estimates and QOF registers. The largest differences appear to be in the South of England. The overall percentage difference between the local estimates and QOF registers is 12%. In general the local estimates are slightly higher than the registered prevalence, as we would expect given the model we developed. The prevalence of GP-registered plus probable/possible cases in our CPRD dataset is about 20% higher than GP-registered prevalence alone, and the average prevalence in our local estimates is 15% higher than aggregated GP registers. Comparing the local estimates with NoAR, which gave a whole population prevalence of exactly 1.00% (66/6593),[1] the estimated prevalence in the East of England is 0.86%, in between the QOF registered prevalence and the NoAR prevalence. This will be explored further in the spatial analysis noted in the original objectives, which could not start until the local estimates were available.

Table 65: comparison of aggregate local estimates with England Regions in derivation dataset

		North East	North West	Yorkshire & Humber	East Midlands	West Midlands	East England	London	South East Coast	South Central	South West	Total
Derivation dataset	**Population**	206,749	1,390,692	509,005	492,177	1,117,862	1,305,431	2,069,328	1,408,348	1,668,509	1,234,539	11,402,640
	Cases	1,466	9,164	3,464	2,675	7,559	7,054	12,824	6,486	6,617	6,750	64,059
	Prevalence	0.71%	0.66%	0.68%	0.54%	0.68%	0.54%	0.62%	0.46%	0.40%	0.55%	0.56%
Local estimates	**Population**	2,276,103	6,195,483	4,596,720	3,933,515	4,844,965	5,133,386	7,515,051	3,926,253	3,665,008	4,648,140	46,734,624
	Cases	19,295	52,148	38,856	33,778	41,815	44,359	54,098	34,705	30,833	41,975	391,862
	Prevalence	0.85%	0.84%	0.85%	0.86%	0.86%	0.86%	0.72%	0.88%	0.84%	0.90%	0.84%
Difference		-14.09%	-18.10%	-16.95%	-31.65%	-18.38%	-31.96%	-10.03%	-41.95%	-44.34%	-35.32%	-27.67%

Table 66: comparison of aggregate local estimates with aggregate QOF registers for England Regions

		North East	North West	Yorks & Humber	East Midlands	West Midlands	East Of England	London	South East Coast	South Central	South West	Total
Local estimates	Population	2,276,103	6,195,483	4,596,720	3,933,515	4,844,965	5,133,386	7,515,051	3,926,253	3,665,008	4,648,140	46,734,624
	Cases	19,295	52,148	38,856	33,778	41,815	44,359	54,098	34,705	30,833	41,975	391,862
	Prevalence	0.85%	0.84%	0.85%	0.86%	0.86%	0.86%	0.72%	0.88%	0.84%	0.90%	0.84%
	Mean	0.84%	0.83%	0.83%	0.86%	0.85%	0.86%	0.73%	0.88%	0.84%	0.92%	
	Median	0.85%	0.83%	0.83%	0.87%	0.85%	0.87%	0.72%	0.88%	0.85%	0.93%	
	IQR	0.14%	0.17%	0.19%	0.16%	0.17%	0.20%	0.20%	0.16%	0.20%	0.20%	
	Minimum	0.24%	0.27%	0.22%	0.24%	0.29%	0.27%	0.26%	0.25%	0.26%	0.23%	
	Maximum	1.25%	2.32%	2.32%	1.34%	1.39%	1.51%	1.22%	1.45%	1.34%	1.35%	
QOF registers	Population	2,245,718	6,084,058	4,526,413	3,877,476	4,763,774	5,052,521	7,384,777	3,864,559	3,611,743	4,570,805	45,981,844
	Cases	19,141	45,741	34,514	28,704	39,267	40,381	39,545	28,644	22,931	36,410	335,278
	Prevalence	0.85%	0.75%	0.76%	0.74%	0.82%	0.80%	0.54%	0.74%	0.64%	0.80%	0.73%
	Mean	0.87%	0.75%	0.78%	0.75%	0.83%	0.81%	0.56%	0.76%	0.63%	0.81%	
	Median	0.81%	0.74%	0.76%	0.76%	0.81%	0.80%	0.53%	0.74%	0.61%	0.80%	
	IQR	0.30%	0.32%	0.30%	0.26%	0.33%	0.31%	0.29%	0.28%	0.24%	0.28%	
	Minimum	0.01%	0.07%	0.02%	0.01%	0.05%	0.03%	0.03%	0.03%	0.04%	0.04%	
	Maximum	2.83%	2.29%	4.41%	1.76%	4.60%	2.57%	2.35%	1.62%	1.99%	1.77%	
Difference		0%	9%	9%	12%	4%	6%	18%	14%	20%	10%	11%

4.8.3 Bland-Altman plots

We externally validated the model-estimated prevalence by carrying out a disagreement analysis between model-estimated and QOF prevalence (%) of RA in practices. We estimated three principal components of disagreement (discordance as measured by Kendall's tau-a, bias as measured by median difference, and calibration as measured by the Theil-Sen median slope). Kendall's tau-a between mean prevalence and prevalence difference of RA at practice level is 0.334 (95% CIs (0.320 – 0.348, p<0.001), showing that prevalence means and model-QOF differences are 33.4 percent more likely to be concordant than to be discordant.. Table 67 shows the percentile differences between practice-level model-estimated and QOF prevalences of RA. The percentile slope of model-estimated prevalence with respect to QOF prevalence of diagnosed RA was 0.1 (95% CI 0.1-0.1). The best way to display the data is to plot the difference between the measurements by the two methods for each subject against their mean, using Bland-Altman plots. Figure 21 shows the Bland-Altman plot for the practice-level QOF and estimated prevalence for RA with no much variation. Figure 22 is a scatter plot of practice-level model-estimated and QOF prevalence of diagnosed RA.

Table 67: Percentile differences between model-estimated and QOF prevalence of Rheumatoid Arthritis

Percent	Percentile	(95% CI)
0	-3.8	-3.8)
25	-0.0	(-0.0, -0.0)
50	0.1	(0.1, 0.1)
75	0.2	(0.2, 0.3)
100	1.0	(1.0, 1.0)

Figure 21: Bland-Altman plot for practice-level model-estimated and QOF prevalence of RA

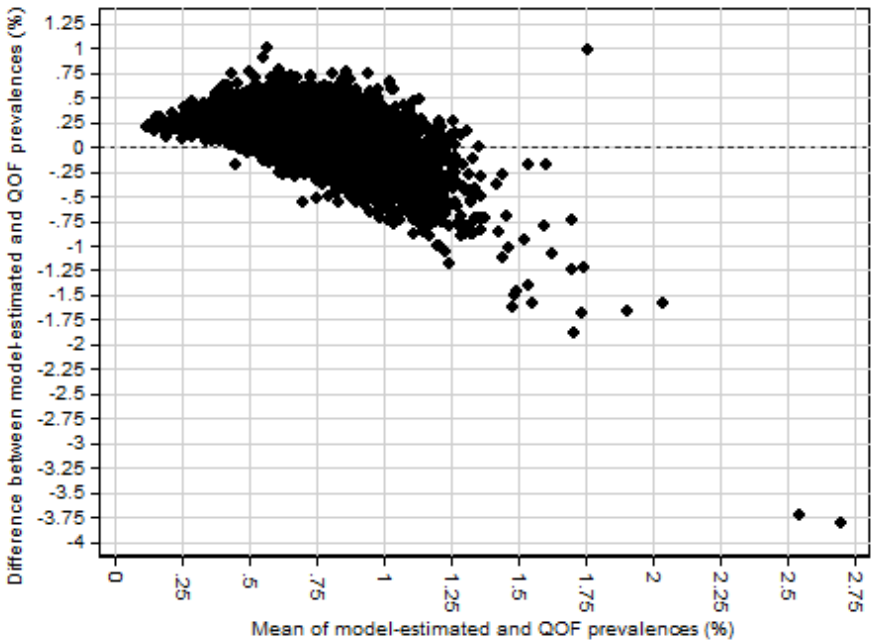
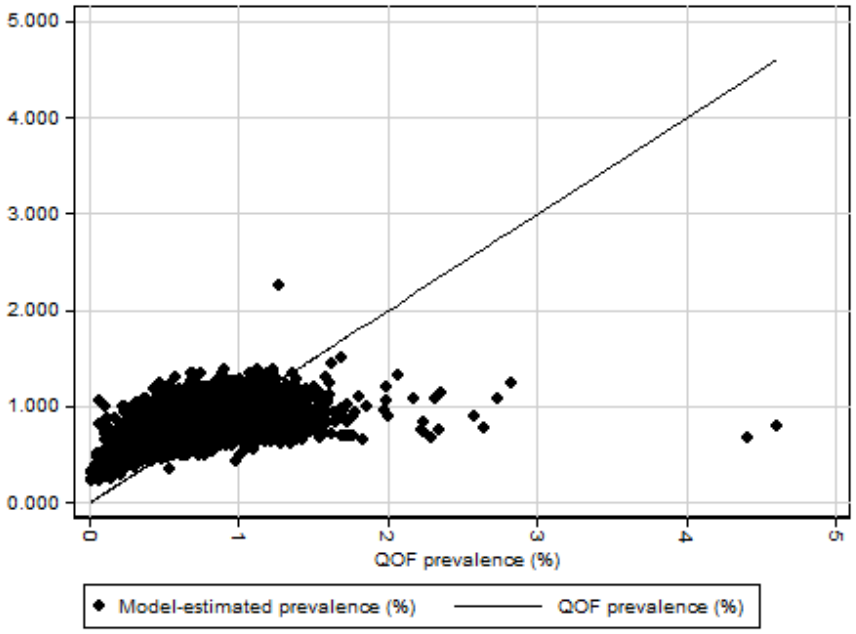


Figure 22: Scatter plot of practice -level model-estimated and QOF prevalence of diagnosed RA



4.9 Production of Scottish local estimates

4.9.1 Methods

We used the model (M8 including doctor diagnosed, algorithm identified and DMARD initiation patients) produced from the CPRD UK RA dataset to estimate Scottish RA prevalence at GP practice and local authority (LA) levels. We compared the prevalences at Health Board levels which were aggregated up from GP practice and LA levels separately. The averages of RA prevalence were compared between Scotland and England at practice level.

4.9.2 Results

Figure 23 and **Figure 24** show the histogram of RA prevalence at practice and LA levels in Scotland.

Figure 23: Histogram of RA prevalence at practice level in Scotland

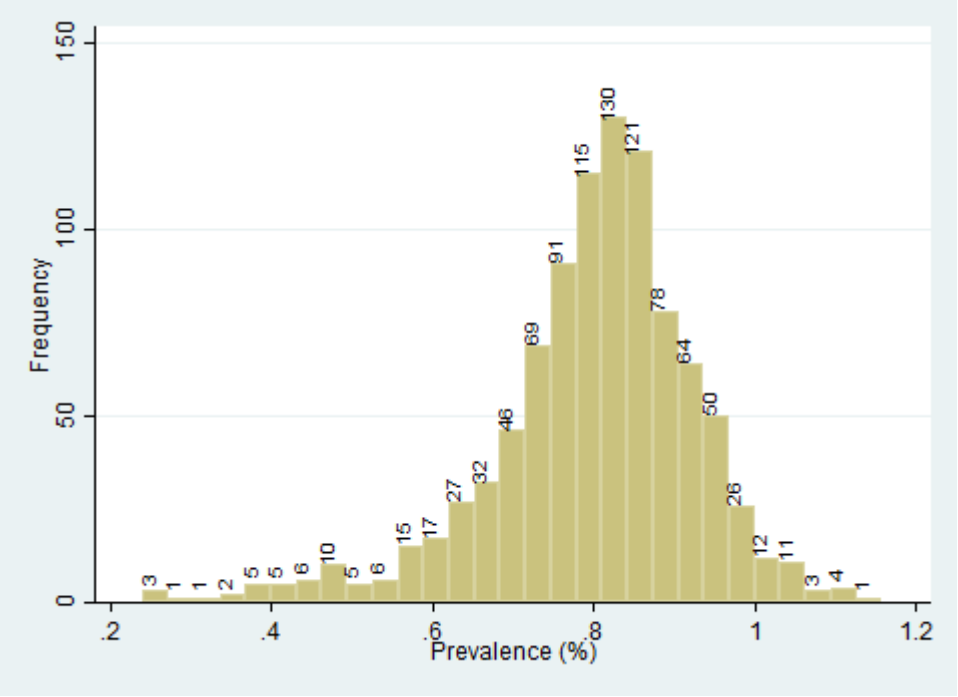


Figure 24: Histogram of RA prevalence at LA level in Scotland

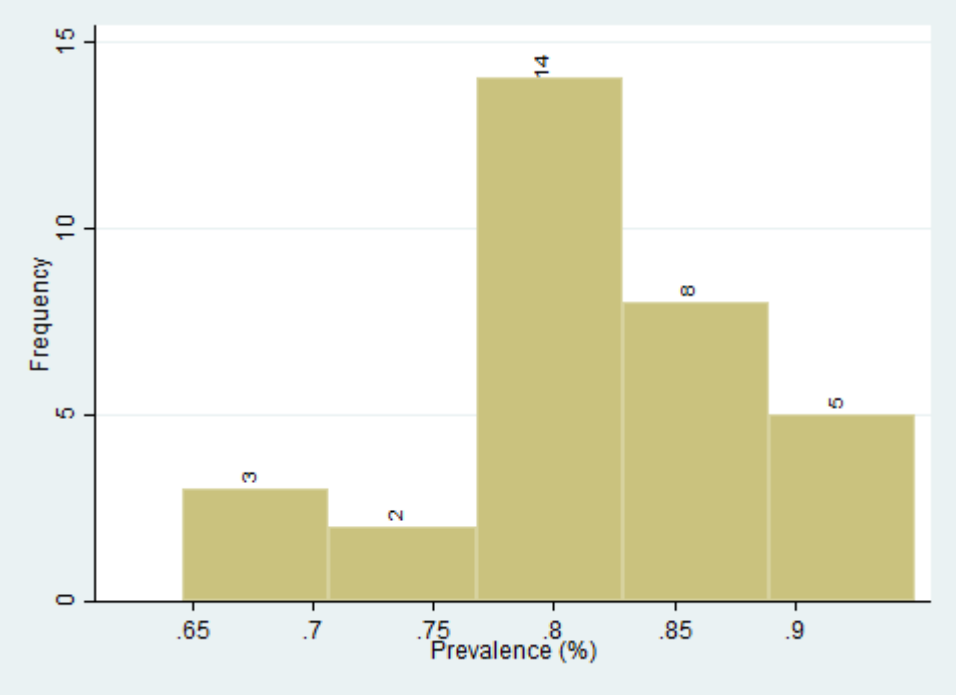


Table 68 shows the prevalence at health board level aggregated up by practice and LA levels separately. There is no much difference between the health board level prevalences from the two levels. But the prevalence at Greater Glasgow & Clyde and Lanarkshire cannot be aggregated up from LA levels due to three LAs fit into two Health Board. **Table 69** shows the average RA prevalence at practice level in Scotland is 0.80% compared to 0.83% in England.

Table 68: Health board level prevalence aggregated up by GP practice and LA levels prevalence

Health Board Code	Health Board Name	Practice population	Practice cases	practice prevalence	LA population	LA cases	LA prevalence
A	Ayrshire & Arran	384390	3463	0.90%	384703	3361.119	0.87%
B	Borders	117029	1035	0.88%	116760	1027.321	0.88%
F	Fife	379131	3177	0.84%	379903	3135.962	0.83%
G	Greater Glasgow & Clyde	1269613	9397	0.74%	N/A	N/A	N/A
H	Highland	325554	2866	0.88%	325237	2795.354	0.86%
L	Lanarkshire	676092	5467	0.81%	N/A	N/A	N/A
N	Grampian	595989	4716	0.79%	595804	4467.081	0.75%
R	Orkney	21099	198	0.94%	21112	190.7207	0.90%
S	Lothian	923251	6458	0.70%	920562	6339.124	0.69%
T	Tayside	425367	3511	0.83%	425214	3525.252	0.83%
V	Forth Valley	316531	2548	0.81%	316244	2545.628	0.80%
W	Western Isles	9748	94	0.96%	26967	256.2058	0.95%
Y	Dumfries & Galloway	154063	1422	0.92%	154115	1416.853	0.92%
Z	Shetland	23045	155	0.67%	23045	189.5186	0.82%

Table 69: Prevalence comparison between Scotland and England at GP practice level.

Prevalence	Scotland	England
Practice Level	0.80%	0.83%

4.9.3 Internal validation

As part of the validation of the ARUK rheumatoid arthritis (RA) prevalence model for the Scottish population to produce estimates at general practice, Scottish Health Board and LA levels, we applied the RA prevalence model developed from UK-wide Clinical Practice Research Datalink (CPRD) data to the Scottish population only.

We dropped patients from other UK countries and fitted a logistic regression model using only Scottish data. The associations between individual risk factor and RA diagnosis were analysed using univariate and multi-variate logistic regression. The associations from Scottish model were also compared with the UK model. The C statistics were calculated using ROC curves.

There are 46,838 Scottish patients in the CPRD UK RA dataset, which makes up 8.64% of the total population. The population counts from different English Regions and UK Countries are shown in **Table 70**. There are 9,255 cases and 37,583 controls (a sample of all controls) in the Scottish dataset. The case: control ratio is 19.24%: 80.76%, which is similar to the UK model (18.79% cases and 81.21% controls).

Table 70: populations in different English Regions or UK Countries CPRD UK dataset

Region/Country	Frequency	Percent	Cum.
North East	8,284	1.53	1.53
North West	57,759	10.65	12.18
Yorkshire & The Humber	19,806	3.65	15.83
East Midlands	20,401	3.76	19.6
West Midlands	44,413	8.19	27.79
East of England	51,748	9.54	37.33
South West	46,271	8.53	45.87
South Central	60,952	11.24	57.11
London	77,612	14.32	71.43
South East Coast	54,529	10.06	81.48
Northern Ireland	13,098	2.42	83.9
Scotland	46,838	8.64	92.54
Wales	40,454	7.46	100
Total	542,165	100	

The baseline characteristics of the Scottish population are shown in **Table 71**. The baseline characteristics of the Scottish population are similar to the UK, with the exception that there were more patients over 75 and less patients from 18-44 age group in the Scottish compared with the UK population. In addition, there were more non-drinkers (36.15%) in the Scottish model compared to the UK model (16.96%).

Table 71: Baseline characteristics of patients involved in the Scottish and UK RA dataset

Predictor variables	Scottish Model	UK Model
Total number of respondents	46,838	455,898
Gender		
Male	21,104 (45.06%)	200,750 (44.03%)
Female	25,734 (54.94%)	255,148 (55.97%)
Total	46,838	455,898
Age group		
18-44	17,240 (36.81%)	227,874 (49.98%)
45-64	13,888 (29.65%)	121,320 (26.61%)
65-74	5,539 (11.83%)	48,289 (10.59%)
>75	10,171 (21.72%)	58,415 (12.81%)
Total	46,838	455,898
Alcohol		
Non-drinker	9,756 (36.15%)	47,026 (16.96%)
Light (<15 units per week)	13,844 (51.30%)	186,443 (67.22%)
Moderate (14-42 units per week)	2,352 (8.71%)	35,249 (12.71%)
Heavy (>42 units per week)	1,034 (3.83%)	8,635 (3.11%)
Total	26,986	277,353
Ethnicity		
White	45,918 (98.04%)	439,563 (96.42%)
Mixed	182 (0.39%)	2,486 (0.55%)
Asian	254 (0.54%)	6,781 (1.49%)
Black	95 (0.20%)	4,092 (0.90%)
Other	389 (0.83%)	2,976 (0.65%)
Total	46,838	455,898
BMI		
Underweight (<18.5)	1,572 (5.11%)	11,645 (3.55%)
Normal (18.5-25)	12,051 (39.17%)	138,274 (42.19%)
Overweight (25-30)	9,791 (31.82%)	105,761 (32.27%)
Obese (>30)	7,352 (23.90%)	72,049 (21.98%)
Total	30,766	327,729
Smoking		
Non-smoker	18,763 (51.93%)	198,552 (54.16%)
Ex-smoker	8,128 (22.50%)	84,474 (23.04%)
Smoker	9,238 (25.57%)	83,562 (22.79%)
Total	36,129	366,588
Deprivation		
1 (least deprived)	11,435 (24.41%)	17,739 (16.03%)
2	6,614 (14.12%)	21,195 (19.15%)
3	9,455 (20.19%)	23,633 (21.35%)
4	9,384 (20.04%)	25,785 (23.30%)
5 (most deprived)	9,949 (21.24%)	22,316 (20.16%)
Total	46,837	110,668

The univariate analysis of individual risk factors is shown in **Table 72**. The associations between individual risk factors and RA diagnoses in the Scottish data were similar to that of the UK. However, light alcohol consumption is a risk factor in the Scottish model but is a protective factor in the UK model. Underweight is a protective factor in the Scottish model while it is a risk factor in the UK model.

Table 73 shows the Scottish multivariate logistic regression model. The odds ratios (ORs) from the Scottish model were similar to the UK model with the exception that moderate alcohol consumption is a risk factor in the Scottish model but is a protective factor in the UK model. Underweight is a protective factor in the Scottish model while it is a risk factor in the UK model.

Table 72: Univariate logistic model for individual risk factors, Scottish & UK models

	Scottish Model			UK model		
Predictor variables	**Odds Ratio**	**P>t**	**[95% Conf. Interval]**	**Odds Ratio**	**P>t**	**[95% Conf. Interval]**
Gender						
Male	1.00			1.00		
Female	2.231	<0.001	[2.125 - 2.343]	2.114	<0.001	[2.083 - 2.146]
Age group						
18-44	1.00			1.00		
45-64	5.120	<0.001	[4.706 - 5.571]	5.189	<0.001	[5.092 - 5.287]
65-74	11.559	<0.001	[10.547 – 12.667]	8.680	<0.001	[8.482 - 8.883]
>75	13.782	<0.001	[12.683- 14.976]	5.289	<0.001	[5.172 - 5.410]
Alcohol						
Non-drinker	1.00			1.00		
Light (<15 units per week)	1.255	0.305	[1.183 – 1.331]	0.906	<0.001	[0.886 – 0.927]
Moderate (14-42 units per week)	0.711	<0.001	[0.630 - 0.803]	0.576	<0.001	[0.557 - 0.596]
Heavy (>42 units per week)	0.567	<0.001	[0.485 - 0.664]	0.532	<0.001	[0.501 - 0.566]
Ethnicity						
White	1.00			1.00		
Mixed	1.799	<0.001	[1.278 - 1.586]	1.497	<0.001	[1.374 - 1.632]
Asian	0.626	0.011	[0.435 - 0.898]	0.965	0.223	[0.910 – 1.022]
Black	0.753	<0.001	[0.572 - 0.990]	0.624	<0.001	[0.573 - 0.679]
Other	0.752	0.042	[0.572 – 0.990]	0.650	<0.001	[0.588 – 0.717]
BMI						
Underweight (<18.5)	0.885	0.111	[0.760 – 1.030]	1.536	<0.001	[1.471 - 1.603]
Normal (18.5-25)	1.00			1.00		
Overweight (25-30)	1.463	<0.001	[1.383 - 1.548]	1.297	<0.001	[1.272 - 1.322]
Obese (>30)	1.618	<0.001	[1.505 - 1.740]	1.720	<0.001	[1.685 - 1.756]

	Scottish Model			UK model		
Smoking						
Non-smoker	1.00			1.00		
Ex-smoker	2.110	<0.001	[1.983 – 2.245]	1.709	<0.001	[1.678 - 1.740]
Smoker	0.861	<0.001	[0.806 – 0.922]	0.773	<0.001	[0.757 – 0.789]
Deprivation						
1 (least deprived)	1.00			1.00		
2	1.765	<0.001	[1.635 - 1.906]	1.082	0.007	[1.022 - 1.145]
3	1.212	<0.001	[1.126 - 1.305]	1.120	<0.001	[1.060 - 1.185]
4	1.274	<0.001	[1.184 - 1.371]	1.225	<0.001	[1.159 - 1.295]
5 (most deprived)	2.016	<0.001	[1.884 - 2.159]	1.312	<0.001	[1.238 - 1.390]

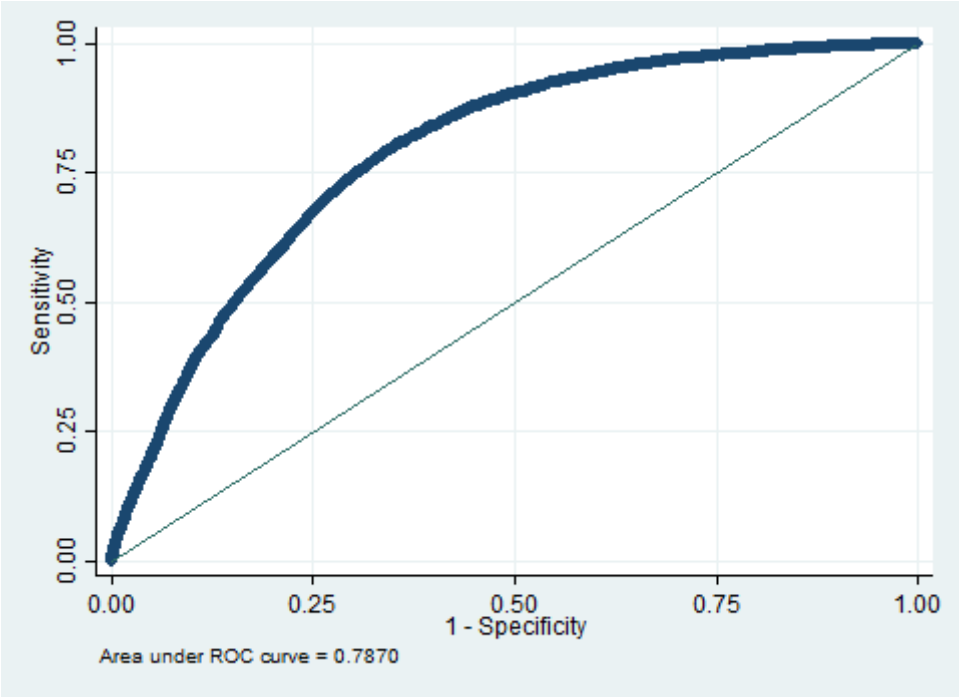
Table 73: Multivariate logistic regression model, Scottish & UK models

	Scottish Model			UK model		
Predictor variables	**Odds Ratio**	**P>t**	**[95% Conf. Interval]**	**Odds Ratio**	**P>t**	**[95% Conf. Interval]**
Gender						
Male	1.00			1.00		
Female	2.214	<0.001	[2.092 - 2.342]	2.207	<0.001	[2.169-2.246]
Age group						
18-44	1.00			1.00		
45-64	4.916	<0.001	[4.508 - 5.362]	4.999	<0.001	[4.897-5.103]
65-74	10.477	<0.001	[9.525 – 11.523]	8.471	<0.001	[8.260-8.687]
>75	12.492	<0.001	[11.459- 13.618]	5.146	<0.001	[5.020-5.275]
Alcohol						
Non-drinker	1.00			1.00		
Light (<15 units per week)	1.353	<0.001	[1.261 - 1.453]	1.065	<0.001	[1.038-1.093]
Moderate (14-42 units per week)	1.027	0.705	[0.891 – 1.183]	0.934	0.003	[0.895-0.976]

	Scottish Model			UK model		
Heavy (>42 units per week)	0.827	0.026	[0.706 - 0.977]	0.894	<0.001	[0.846-0.944]
Ethnicity						
White	1.00			1.00		
Mixed	1.734	0.003	[1.234 – 2.489]	1.561	<0.001	[1.413-1.723]
Asian	1.397	0.096	[0.942 - 2.072]	1.599	<0.001	[1.496-1.709]
Black	0.824	0.647	[0.367 – 1.884]	0.896	0.027	[0.813-0.988]
Other	1.196	0.246	[0.884 - 1.617]	1.010	0.856	[0.904-1.129]
BMI						
Underweight (<18.5)	0.980	0.803	[0.833 - 1.152]	1.194	<0.001	[1.142-1.248]
Normal (18.5-25)	1.00			1.00		
Overweight (25-30)	1.187	<0.001	[1.116 - 1.263]	1.175	<0.001	[1.149-1.202]
Obese (>30)	1.299	<0.001	[1.290 - 1.396]	1.476	<0.001	[1.443-1.509]
Smoking						
Non-smoker	1.00			1.00		
Ex-smoker	1.542	<0.001	[1.447 - 1.644]	1.526	<0.001	[1.483-1.571]
Smoker	1.017	0.691	[0.935 - 1.105]	1.095	<0.001	[1.062-1.128]
Deprivation						
1 (least deprived)	1.00			1.00		
2	1.123	<0.001	[1.085 - 1.162]	1.117	0.032	[1.014-1.230]
3	1.130	<0.001	[1.016 - 1.257]	1.124	0.017	[1.031-1.227]
4	1.272	<0.001	[1.133 - 1.428]	1.244	0.002	[1.130-1.370]
5 (most deprived)	1.374	<0.001	[1.291 - 1.462]	1.372	<0.001	[1.261-1.494]
_cons	0.312	<0.001	[0.029 - 0.034]	0.035	<0.001	[0.032-0.037]

The Scottish model discrimination as measured by the ROC curve is shown in Error! Reference source n ot found. (c statistic 0.7870). This measures the same model, M6, used in the UK model i.e. logistic regression model including patients with CPRD doctor diagnosed RA, HES RA diagnosis, algorithm identified RA cases and DMARDs cases, but in the case of the UK dataset the c statistic was 0.7599 [0.759 - 0.761]. So the Scottish model, with a much smaller population, discriminates somewhat better than the UK model.

Figure 25: ROC curve of Scottish model



The reasons for the better discrimination in the Scottish data are evident in the multivariable model with higher ORs in Scotland for age >75 and in the previously noted alcohol variable. However the variables included and the direction of the effects are similar, so it is likely that applying Scotland-specific ORs to the local risk factor data will have a fairly small impact on the local estimates.

4.10 Production of Wales local estimates

4.10.1 Methods

A major stumbling block in producing estimates for Wales during the main project was the lack of practice/MSOA lookup tables, so that it was not possible to convert from one geography to the other. These tables were not produced until late 2017, so Wales estimates were finally produced in 2018. As shown in Figure 1, the representation of CPRD practices in Wales is relatively low, comprising only about 40 practices. We therefore decided to apply the model developed using UK CPRD data to the Wales population, as we did for back pain and osteoarthritis. We used the RA model M8 i.e. including doctor diagnosed, algorithm identified and DMARD initiation patients).

As shown in Table 74, there were significant differences between the variables in the UK model and the local Wales data. It proved to be impossible to match alcohol consumption data between the Health Survey for England and Welsh Health Survey lifestyle trends (2015) categories, as the latter included the categories drinking above guidelines on a day in the past week, heavy (binge) drinking and very heavy drinking. We therefore dropped alcohol from the England model and local Wales risk

factor data. Therefore the final model included only age, gender, BMI (four categories), smoking (three categories), deprivation (fifths) and ethnicity (five categories).

Table 74: choice of Wales local risk factor data

Geography 1	Risk factor	Definition	Action in national regression model	Geography2	Name of source
Health Board/ practice	Smoking	As the Welsh Health Survey lifestyle trends (2015) uses current smokers + ecigarette users only, we used QOF data	Same	LA, MSOA	QOF data (smoker/ex-smoker/non-smoker)
Practice/ health board	Age & gender	Same as England	Same	LA, MSOA	Population estimates by middle layer super output area and age group
LA	Alcohol consumption	Heavy drinking & binge drinking,		National level data	Welsh Health Survey lifestyle trends (2015)
LA	Ethnicity	Same	Same	Council area	Ethnic Group Demographics
LA	BMI	Overweight or obese and obese only (so overweight only can be calculated, giving three categories), we used 4 categories	Combine underweight and normal range to obtain 3 categories	National level data	Welsh Health Survey lifestyle trends (2015)
LSOA	Deprivation	Rank	Same		Welsh Index of Multiple Deprivation

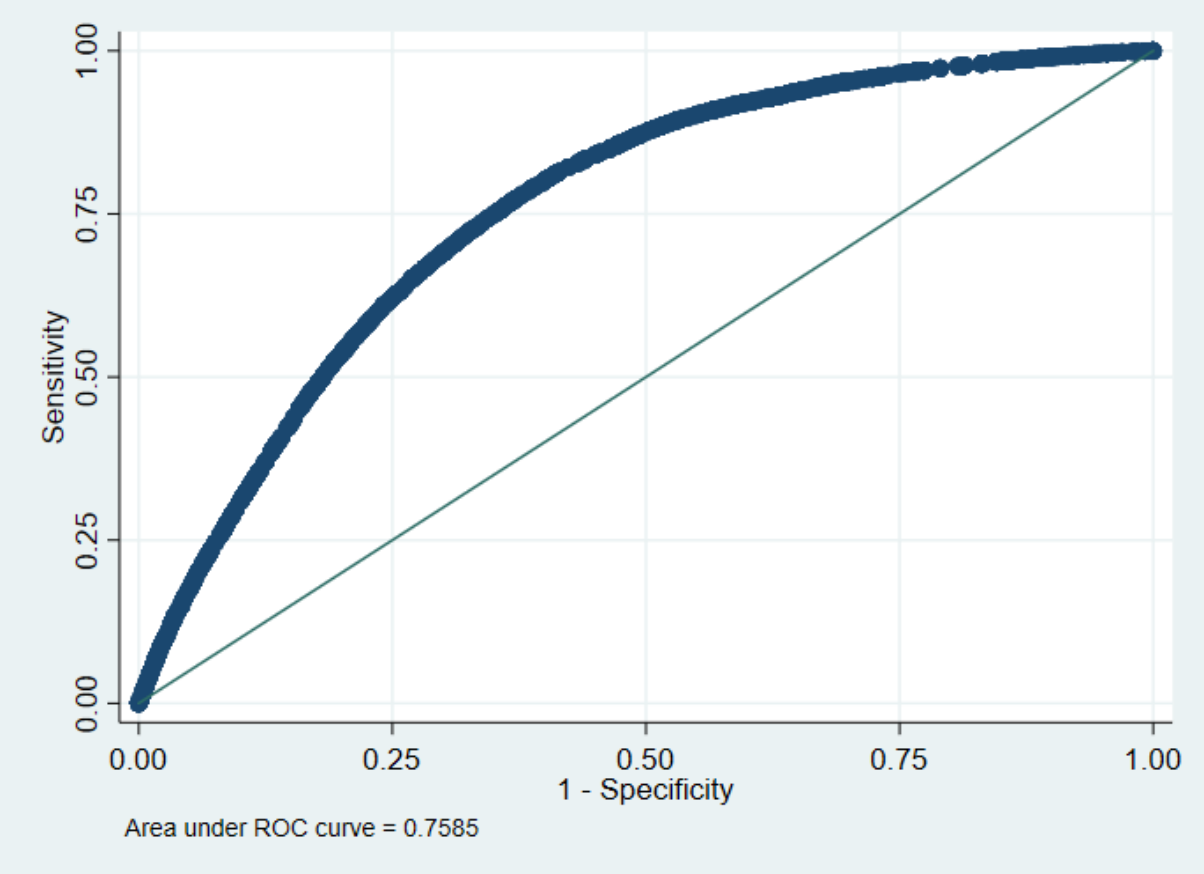
The averages of RA prevalence were compared between Wales and England at practice level.

External validation was carried out in the same way as for England, using the three principal parameters of the Bland-Altman plot. These parameters are the Kendall's tau-a between estimated and registered prevalence, the mean sign of the difference between estimated and registered prevalences, and the Kendall's tau-a between the mean of the two prevalences and the difference between the two prevalences. These measure the three principal components of disagreement between two measurements of the same thing, which are discordance, bias and scale discrepancy, respectively.

4.10.2 Results

Figure 26 shows the discrimination (ROC curve) of the Wales regression model. The C-statistic is 0.76, which is lower than the 0.79 of the Scottish model fitted directly from Scottish data, but it still performs well given the decreased number of variables included.

Figure 26: Discrimination (ROC curve) of Wales regression model



As for the England estimates, we carried out an external validation against QOF-registered data. Figure 27 shows Bland-Altman plot for practice-level model-estimated and QOF prevalence of RA. We see, visually, that the correlation between the two prevalences is not too strong, that the estimated prevalence is usually greater than the registered prevalence by about 0.25%, and that the cloud of data points in the Bland-Altman plot tends to slope downwards (indicating that the scale of variability is smaller for estimated prevalences than for registered prevalence).

The Kendall's tau-a between estimated and registered parameters is 0.221 (95% CI, 0.158 to 0.282). This indicates that there is clearly a correlation between the two measures in the population of Welsh practices, but that this correlation isn't very strong, because a pair of practices is only 15.8 percent to 28.2 percent more likely to be concordant than to be discordant. (A pair of practices is concordant if the one with the higher estimated prevalence also has the higher registered prevalence, and discordant if the one with the higher estimated prevalence has the lower registered prevalence.)
The mean sign of the difference between the estimated prevalence and the registered prevalence is 0.879 (95% CI, 0.803 to 0.943), indicating that the estimated-registered prevalence difference is 80.3 percent to 94.3 percent more likely to be positive than to be negative. This indicates that estimated prevalences are positively biassed as an estimate of registered prevalences. And the Kendall's tau-a between the mean prevalence and the difference between estimated and registered prevalences is -0.230 (95% CI, -0.295 to -0.162), indicating that the absolute difference between 2 estimated

prevalences is 16.2 percent to 29.5 percent less likely to be greater than the absolute difference between the two corresponding registered prevalences than to be less than the absolute difference between the two corresponding registered prevalences. This indicates that the scale of variability between estimated prevalences is smaller than the scale of differences between registered prevalences.

Figure 27: Bland-Altman plot for practice-level model-estimated and QOF prevalence of RA

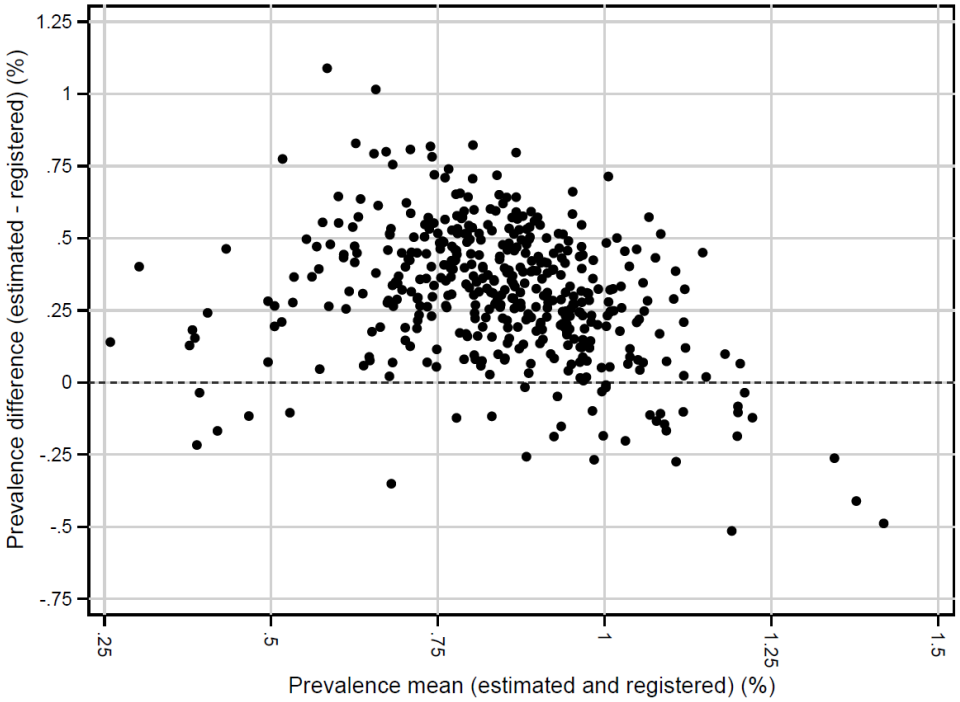


Figure 28 is a scatter plot of estimated prevalences against registered prevalences (with a 45 degree line of equality).

Figure 28: scatter plot of practice -level model-estimated and QOF prevalence of diagnosed RA

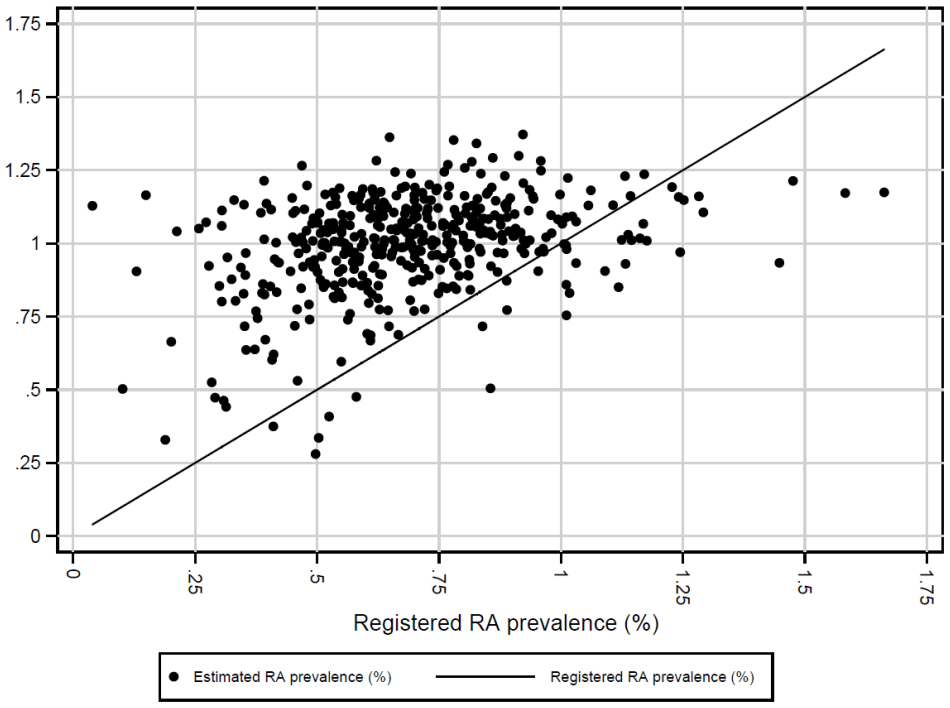
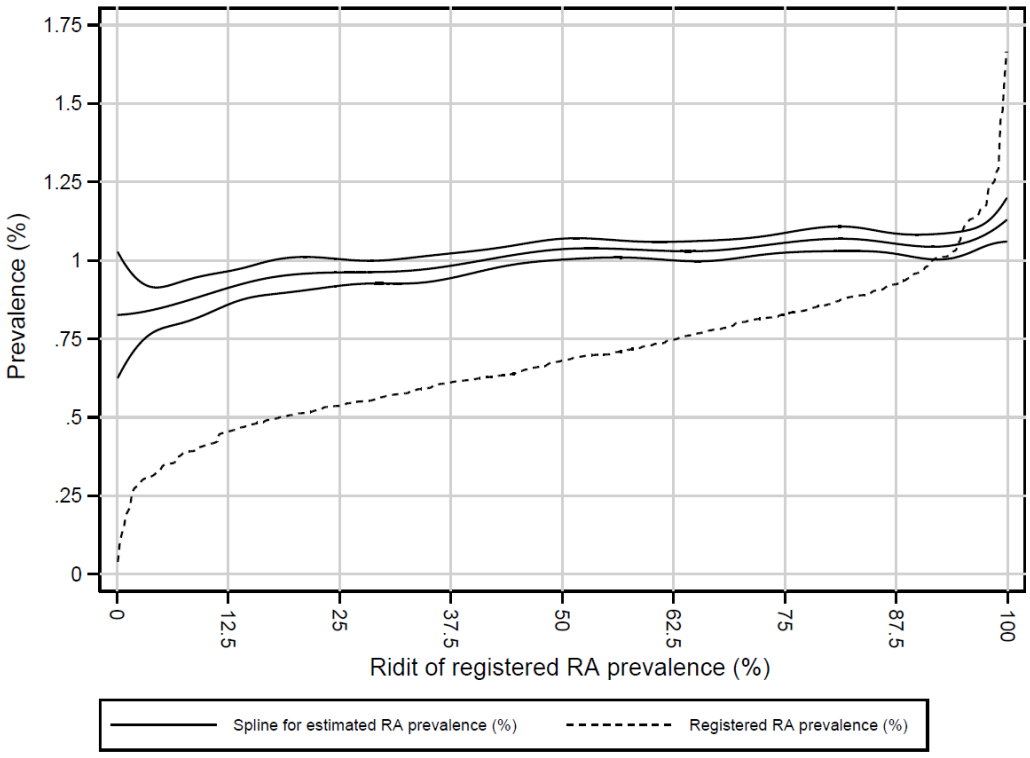


Figure 29 shows a cubic ridit spline calibration curve of estimated prevalence with respect to registered prevalence. Ridits are like ranks, only expressed on a scale from 0 to 100 percent instead of from 1 to N. (So ridits, unlike ranks, are not affected by sample number.) A ridit spline is a spline in the ridit of registered prevalence. Ridit splines are discussed in Newson (2017). In this case, we have used regression to fit a cubic spline of estimated prevalence with respect to the ridit of registered prevalence, with 95% confidence intervals for the value of the spline at each ridit. We have also plotted the registered prevalence against its ridit, giving the percentile registered prevalence continuously for each percent on the ridit scale. We see, once again, that estimated prevalences are expected to be higher than the corresponding registered prevalence, over most of the range of registered prevalences, but that estimated prevalences are expected to be lower than the corresponding registered prevalence if the corresponding registered prevalence is in the highest 5 percent of registered prevalences.

Figure 29: cubic ridit spline calibration curve of estimated prevalence with respect to registered prevalence



5 References

1. Symmons D, Turner G, Webb R, Asten P, Barrett E, et al. The prevalence of rheumatoid arthritis in the United Kingdom: new estimates for a new century. Rheumatology 2002;41(7):793-800. doi: 10.1093/rheumatology/41.7.793 Link: http://rheumatology.oxfordjournals.org/content/41/7/793.abstract.

2. Symmons DPM, Bankhead CR, Harrison BJ, Brennan P, Silman AJ, et al. Blood transfusion, smoking, and obesity as risk factors for the development of rheumatoid arthritis. Results from a primary care-based incident case-control study in Norfolk, England. Arthritis & Rheumatism 1997;40:1955-61. Link: http://onlinelibrary.wiley.com/doi/10.1002/art.1780401106/abstractfiles/718/abstract.html.

3. Maxwell JR, Gowers IR, Moore DJ, Wilson AG. Alcohol consumption is inversely associated with risk and severity of rheumatoid arthritis. Rheumatology (Oxford) 2010;49(11):2140-6. doi: 10.1093/rheumatology/keq202 Link: http://www.ncbi.nlm.nih.gov/pubmed/20667949.

4. Hutchinson D, Shepstone L, Moots R, Lear JT, Lynch MP. Heavy cigarette smoking is strongly associated with rheumatoid arthritis (RA), particularly in patients without a family history of RA. Annals of the Rheumatic Diseases 2001;60:223-27. Link: http://ard.bmj.com/content/60/3/223.shortfiles/723/223.html.

5. Bergstrom U, Jacobsson LT, Nilsson JA, Wirfalt E, Turesson C. Smoking, low formal level of education, alcohol consumption, and the risk of rheumatoid arthritis. Scandinavian journal of rheumatology 2013;42(2):123-30. doi: 10.3109/03009742.2012.723744 Link: http://www.ncbi.nlm.nih.gov/pubmed/23126587.

6. Bengtsson C, Nordmark B, Klareskog L, Lundberg I, Alfredsson L. Socioeconomic status and the risk of developing rheumatoid arthritis: results from the Swedish EIRA study. Annals of the Rheumatic Diseases 2005;64:1588-94. Link: http://ard.bmj.com/content/64/11/1588.shortfiles/813/1588.html.

7. Stolt P, Källberg H, Lundberg I, Sjögren B, Klareskog L, et al. Silica exposure is associated with increased risk of developing rheumatoid arthritis: results from the Swedish EIRA study. Annals of the Rheumatic Diseases 2005;64:582-86. Link: http://ard.bmj.com/content/64/4/582.shortfiles/734/582.htmlfiles/733/582.html.

8. Hazes JM, Dijkmans BA, Vandenbroucke JP, de Vries RR, Cats A. Lifestyle and the risk of rheumatoid arthritis: cigarette smoking and alcohol consumption. Ann Rheum Dis 1990;49(12):980-2. Link: http://www.ncbi.nlm.nih.gov/pubmed/2270970.

9. Jin Z, Xiang C, Cai Q, Wei X, He J. Alcohol consumption as a preventive factor for developing rheumatoid arthritis: a dose-response meta-analysis of prospective studies. Ann Rheum Dis 2013 doi: 10.1136/annrheumdis-2013-203323 Link: http://www.ncbi.nlm.nih.gov/pubmed/23897767.

10. Kallberg H, Jacobsen S, Bengtsson C, Pedersen M, Padyukov L, et al. Alcohol consumption is associated with decreased risk of rheumatoid arthritis: results from two Scandinavian case-control studies. Ann Rheum Dis 2009;68(2):222-7. doi: 10.1136/ard.2007.086314 Link: http://www.ncbi.nlm.nih.gov/pubmed/18535114.

11. Voigt LF, Koepsell TD, Nelson JL, Dugowson CE, Daling JR. Smoking, obesity, alcohol consumption, and the risk of rheumatoid arthritis. Epidemiology 1994;5(5):525-32. Link: http://www.ncbi.nlm.nih.gov/pubmed/7986867.

12. Li X, Sundquist J, Sundquist K. Socioeconomic and occupational risk factors for rheumatoid arthritis: a nationwide study based on hospitalizations in Sweden. J Rheumatol 2008;35(6):986-91. Link: http://www.ncbi.nlm.nih.gov/pubmed/18464310.

13. Pedersen M, Jacobsen S, Klarlund M, Frisch M. Socioeconomic status and risk of rheumatoid arthritis: a Danish case-control study. J Rheumatol 2006;33(6):1069-74. Link: http://www.ncbi.nlm.nih.gov/pubmed/16622905.

14. Di Giuseppe D, Orsini N, Alfredsson L, Askling J, Wolk A. Cigarette smoking and smoking cessation in relation to risk of rheumatoid arthritis in women. Arthritis Res Ther 2013;15(2):R56. doi: 10.1186/ar4218 Link: http://www.ncbi.nlm.nih.gov/pubmed/23607815.

15. Criswell LA, Merlino LA, Cerhan JR, Mikuls TR, Mudano AS, et al. Cigarette smoking and the risk of rheumatoid arthritis among postmenopausal women:: Results from the Iowa Women's Health Study. The American journal of medicine 2002;112:465-71. Link: http://www.sciencedirect.com/science/article/pii/S0002934302010513files/731/S0002934302010513.html.

16. Stolt P, Bengtsson C, Nordmark B, Lindblad S, Lundberg I, et al. Quantification of the influence of cigarette smoking on rheumatoid arthritis: results from a population based case-control study, using incident cases. Annals of the Rheumatic Diseases 2003;62:835-41. Link: http://ard.bmj.com/content/62/9/835.shortfiles/726/835.htmlfiles/725/835.html.

17. Sugiyama D, Nishimura K, Tamaki K, Tsuji G, Nakazawa T, et al. Impact of smoking as a risk factor for developing rheumatoid arthritis: a meta-analysis of observational studies. Annals of the Rheumatic Diseases 2010;69:70-81. Link: http://ard.bmj.com/content/69/01/70.shortfiles/729/70.html.

18. Heliövaara M, Aho K, Knekt P, Impivaara O, Reunanen A, et al. Coffee consumption, rheumatoid factor, and the risk of rheumatoid arthritis. Annals of the Rheumatic Diseases 2000;59:631-35. Link: http://ard.bmj.com/content/59/8/631.shortfiles/741/631.html.

19. Mikuls TR, Cerhan JR, Criswell LA, Merlino L, Mudano AS, et al. Coffee, tea, and caffeine consumption and risk of rheumatoid arthritis: results from the Iowa Women's Health Study. Arthritis & Rheumatism 2002;46:83-91. Link: http://onlinelibrary.wiley.com/doi/10.1002/1529-0131(200201)46:1%3C83::AID-ART10042%3E3.0.CO;2-D/fullfiles/743/full.html.

20. Alamanos Y, Voulgari PV, Drosos AA. Incidence and Prevalence of Rheumatoid Arthritis, Based on the 1987 American College of Rheumatology Criteria: A Systematic Review. Seminars in Arthritis and Rheumatism 2006;36:182-88. doi: 10.1016/j.semarthrit.2006.08.006 Link: http://www.sciencedirect.com/science/article/pii/S0049017206001107.

21. Symmons DPM, Silman AJ. The Norfolk Arthritis Register (NOAR). Clinical and experimental rheumatology 2003;21(Supplement 31):5. Link: http://www.clinexprheumatol.org/abstract.asp?a=2197.

22. Rahman N PE, Bhatia K. Arthritis and musculoskeletal conditions in Australia 2005: with a focus on osteoarthritis, rheumatoid arthritis and osteoporosis: AIHW, 2005 Link: http://www.aihw.gov.au/publication-detail/?id=6442467774.

23. Formica MK, McAlindon TE, Lash TL, Demissie S, Rosenberg L. Validity of self-reported rheumatoid arthritis in a large cohort: results from the Black Women's Health Study. Arthritis Care Res (Hoboken) 2010;62(2):235-41. doi: 10.1002/acr.20073 Link: http://onlinelibrary.wiley.com/doi/10.1002/acr.20073/full.

24. Star VL, Scott JC, Sherwin R, Lane N, Nevitt MC, et al. Validity of self-reported rheumatoid arthritis in elderly women. J Rheumatol 1996;23(11):1862-5. Link: http://www.ncbi.nlm.nih.gov/pubmed/8923357.

25. Kvien TK, Glennas A, Knudsrod OG, Smedstad LM. The validity of self-reported diagnosis of rheumatoid arthritis: results from a population survey followed by clinical examinations. J Rheumatol 1996;23(11):1866-71. Link: http://www.ncbi.nlm.nih.gov/pubmed/8923358.

26. Aletaha D, Neogi T, Silman AJ, Funovits J, Felson DT, et al. 2010 Rheumatoid arthritis classification criteria: an American College of Rheumatology/European League Against Rheumatism collaborative initiative. Annals of the Rheumatic Diseases 2010;69(9):1580-88. doi: 10.1136/ard.2010.138461 Link: http://ard.bmj.com/content/69/9/1580.abstract.

27. Department of Health. General Practice Physical Activity Questionnaire (GPPAQ). In: Health Do, ed. Online, 2014 Link: https://www.gov.uk/government/publications/general-practice-physical-activity-questionnaire-gppaq.

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

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

30. Ezzati M, Vander Hoorn S, Rodgers A, Lopez AD, Mathers CD, et al. Estimates of global and regional potential health gains from reducing muliple major risk factors. The Lancet 2003;362(9380):271-80. doi: http://dx.doi.org/10.1016/S0140-6736(03)13968-2 Link: http://www.sciencedirect.com/science/article/pii/S0140673603139682.

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

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

33. Symmons D, Turner G, Webb R, Asten P, Barrett E, et al. The prevalence of rheumatoid arthritis in the United Kingdom: new estimates for a new century. Rheumatology 2002;41(7):793-800. Link: http://www.ncbi.nlm.nih.gov/pubmed/12096230.

34. Raza K, Stack R, Kumar K, Filer A, Detert J, et al. Delays in assessment of patients with rheumatoid arthritis: variations across Europe. Annals of the Rheumatic Diseases 2011;70(10):1822-25. doi: 10.1136/ard.2011.151902 Link: http://ard.bmj.com/content/70/10/1822.abstract.

35. van der Linden MPM, le Cessie S, Raza K, van der Woude D, Knevel R, et al. Long-term impact of delay in assessment of patients with early arthritis. Arthritis & Rheumatism 2010;62(12):3537-46. doi: 10.1002/art.27692 Link: http://dx.doi.org/10.1002/art.27692.

36. Smolen JS, Breedveld FC, Burmester GR, Bykerk V, Dougados M, et al. Treating rheumatoid arthritis to target: 2014 update of the recommendations of an international task force. Annals of the Rheumatic Diseases 2015 doi: 10.1136/annrheumdis-2015-207524 Link: http://ard.bmj.com/content/early/2015/05/12/annrheumdis-2015-207524.abstract.

37. Raza K, Filer A. The therapeutic window of opportunity in rheumatoid arthritis: does it ever close? Annals of the Rheumatic Diseases 2015;74(5):793-94. doi: 10.1136/annrheumdis-2014-206993 Link: http://ard.bmj.com/content/74/5/793.short.

6 Appendix: additional information

6.1 ELSA outcome and risk factor definitions

Wave 2

Pos. = 523-4 **Variable =** HeArt1-2 **Variable label =** Type of arthritis (1st mention)
- Value = 1 Label = Osteoarthritis?
- Value = 2 Label = Rheumatoid arthritis?
- Value = 3 Label = Some other kind of arthritis?

Pos. = 6604-6 **Variable =** bheart1-3 **Variable label =** Type of arthritis reported at Wave 1 (1st mention)
- Value = 1 Label = ... osteoarthritis?
- Value = 2 Label = ... rheumatoid arthritis?
- Value = 3 Label = ... some other kind of arthritis?

Wave 3

Pos. = 507 **Variable =** heartra **Variable label =** Whether has rheumatoid arthritis
- Value = 0 Label = Not mentioned
- Value = 1 Label = Mentioned

Pos. = 459 **Variable =** hedbwar **Variable label =** Chronic: diagnosed arthritis fed forward
- Value = 1 Label = Chronic lung disease such as chronic bronchitis or emphysema
- Value = 2 Label = Asthma
- Value = 3 Label = Arthritis (including osteoarthritis, or rheumatism)
- Value = 4 Label = Osteoporosis, sometimes called thin or brittle bones
- Value = 5 Label = Cancer or a malignant tumour (excluding minor skin cancers)
- Value = 6 Label = Parkinson s disease
- Value = 7 Label = Any emotional, nervous or psychiatric problems
- Value = 8 Label = Alzheimer s disease
- Value = 9 Label = Dementia, senility or another serious memory impairment

Pos. = 468 **Variable =** hedbmar **Variable label =** Reason disputed arthritis diagnosis fed forward
- Value = 1 Label = Never diagnosed
- Value = 2 Label = No longer has
- Value = 3 Label = Did not have previously, but has now
- Value = 4 Label = Misdiagnosed

Pos. = 477 **Variable =** hedbdar **Variable label =** Whether confirms arthritis diagnosis
- Value = 1 Label = Yes
- Value = 2 Label = No
- Value = 3 Label = Not read out as didn t make sense

Pos. = 486 **Variable =** hedbsar **Variable label =** Whether still has arthritis
- Value = 1 Label = Yes
- Value = 2 Label = No

Pos. = 493 **Variable =** dhedibar **Variable label =** Chronic: arthritis diagnosis newly reported
- Value = 0 Label = Not mentioned
- Value = 1 Label = Mentioned

Wave 4

Pos. = 649 **Variable =** heartra **Variable label =** Whether has rheumatoid arthritis
- Value = 0 Label = Not mentioned
- Value = 1 Label = Mentioned

Wave 5

Pos. = 650 **Variable =** heartra **Variable label =** Whether has rheumatoid arthritis
- Value = 0 Label = Not mentioned
- Value = 1 Label = Mentioned

6.1.1 ELSA RA outcome creation process

Wave 0

Based on the previous section a variable **'w0heartra'** was created at Wave 0 based on the information from variables illsm1-6 (Wave 0 1998 and 1999) and discode1-3 (Wave 0 2001). At Wave0 (1998 and 1999) information is coded in w0illsm1-6 with a label 34 -"Arthritis/rheumatism/fibrositis". Therefore, Wave 0 1998 and 1999 will be excluded.
w0heartra "Whether has rheumatoid arthritis" was equal to:

- 0 "Not mentioned" (w0discode1-3 had any label but NOT 51)
- 1 "Mentioned" (w0discode1-3 had a label of 51).

w0discode1-3 label 51 is "Rheumatoid arthritis", which is appropriate. This condition identified 38 RA cases.

Wave 1
w1heartra variable "Whether has rheumatoid arthritis" was created based on the answers to questions heart1, heart2 and heart3.
W1heartra was equal to:

- 0 "Not mentioned" (if heart1-3 had a label either -1 OR 1 OR 3)
- 1 "Mentioned" (if heart1-3 had a label of 2)
- Missing (if heart1-3 had a label of -9 OR -8)

Respondents were given -1 "Not applicable" for questions heart1-3 if they previously answered "No" to a question HeDiab ("If ever had an arthritis diagnosis").

Table 75: RA cases at Wave 1

Whether has RA	Frequency	Percentage
Not mentioned	10,825	89.47%
Mentioned	835	6.90%
Missing	439	3.63%
Total	12,099	100%

Wave 2
W2heartra variable "Whether has rheumatoid arthritis" was created based on the answers to questions HeArt1 and HeArt2.
W2heartra was equal to:

- 0 "Not mentioned" (if HeArt1-2 had a label either -1 OR 1 OR 3)
- 1 "Mentioned" (if HeArt1-2 had a label of 2)
- Missing (if HeArt1-2 had a label of -9 OR -8)

Respondents were given -1 "Not applicable" for questions HeArt1 and HeArt2 if they previously answered "No" to a question HeDiab ("If ever had an arthritis diagnosis").

Table 76 RA cases at Wave 2

Whether has RA	Frequency	Percentage
Not mentioned	9,216	97.71%
Mentioned	119	1.26%
Missing	97	1.03%
Total	9,432	100.00%

At Wave 2 there is a question bheart1-3 ("Type of arthritis reported at Wave 1 (1^{st}, 2^{nd}, 3^{rd} mention)"). Therefore, it was checked whether there was an overlap between new cases at Wave 2 (coded at w2heartra) and old cases at Wave 1 using this variable. Two cases were overlapping, therefore, these two cases were coded as old RA cases (instead of 119 it is 117).

Wave 3

W3heartra variable "Whether has rheumatoid arthritis" was created based on the answers to question heartra.

W3heartra was equal to:

- 0 "Not mentioned" (if heartra had a label either -1 OR 0)
- 1 "Mentioned" (if heartra had a label of 1)
- Missing (if heartra had a label of -9 OR -8)

Respondents were given -1 "Not applicable" for question heartra if they previously answered "No" to a question HeDiab ("If ever had an arthritis diagnosis").

Table 77 RA cases at Wave 3

Whether has RA	Frequency	Percentage
Not mentioned	9,142	93.56%
Mentioned	629	6.44%
Missing	0	0%
Total	9,771	100.00%

At Wave 3 there is dhedibar question asking 'Chronic: arthritis diagnosis newly reported". There are 676 (6.92%) cases out of 9,771. Therfore it seems that 629 out of 676 are RA cases as it is confirmed in the heartra question.

Wave 4

W4heartra variable "Whether has rheumatoid arthritis" was created based on the answers to question heartra.

W4heartra was equal to:

- 0 "Not mentioned" (if heartra had a label either -1 OR 0)
- 1 "Mentioned" (if heartra had a label of 1)
- Missing (if heartra had a label of -9 OR -8)

Respondents were given -1 "Not applicable" for question heartra if they previously answered "No" to a question HeDiab ("If ever had an arthritis diagnosis").

Table 78 RA cases at Wave 4

Whether has RA	Frequency	Percentage
Not mentioned	9,966	90.19%
Mentioned	704	6.37%
Missing	380	3.44%
Total	11,050	100%

Wave 5

W5heartra variable "Whether has rheumatoid arthritis" was created based on the answers to question heartra.

W5heartra was equal to:

- 0 "Not mentioned" (if heartra had a label either -1 OR 0)
- 1 "Mentioned" (if heartra had a label of 1)
- Missing (if heartra had a label of -9 OR -8)

Respondents were given -1 "Not applicable" for question heartra if they previously answered "No" to a question HeDiab ("If ever had an arthritis diagnosis").

Table 79 RA cases at Wave 5

Whether has RA	Frequency	Percentage
Not mentioned	9,295	90.47%
Mentioned	625	6.08%
Missing	354	3.45%
Total	10,274	100%

6.1.2 ELSA risk factor questions

Table 80 below contains all ELSA variables related to risk factors identified in the literature search. However, only the highlighted questions will be considered as they are more appropriate for our analysis purpose. Usual coding for ELSA variables with missing data is as follows:

Value = -9 Label = No answer/refused
Value = -8 Label = Don't know
Value = -6 Label = Schedule not obtained
Value = -2 Label = Schedule not applicable
Value = -1 Label = Item not applicable

All other coding of all variable categories is available on request. Variables used are listed below.

Table 80: ELSA risk factor variables

Wave	Variable position	Variable name	Variable definition
Wave 0 1998	524	drating	(D) Total Units of alcohol/week
Wave 0 1998	528	overlim	(D) Drinking in relation to weekly limits
Wave 0 1998	541	dnoft	Frequency drank any alcoholic drink last 12 mths
Wave 0 1998	612	dnnow	Whether drink nowadays
Wave 0 1998	613	dnany	Whether drinks occasionally or never drinks
Wave 0 1998	614	dnevr	Whether always non-drinker
Wave 0 1998	89	topqual2	(D) Highest Educational Qualification - Students separate
Wave 0 1998	90	topqual3	(D) Highest Educational Qualification
Wave 0 1998	38	sex	Sex
Wave 0 1998	249	bmival	(D) Valid BMI - inc estimated>130kg
Wave 0 1998	250	bmi	(D) BMI - inc unreliable measurements
Wave 0 1998	251	bmivg4	(D) Valid BMI (grouped:<20,20-25,25-30,30+)
Wave 0 1998	252	bmivg6	(D) Valid BMI (grouped:<20,20-25,25-30,30-35,35-40,40+)
Wave 0 1998	102	schoh	(D) Social Class of HOH - Harmonised
Wave 0 1998	103	schohg7	(D) Social Class of HOH - I,II,IIIN,IIIM,IV,V,Others
Wave 0 1998	104	schohg6	(D) Social Class of HOH - I,II,IIIN,IIIM,IV,V
Wave 0 1998	105	schohg4	(D) Social Class of HOH: I/II,IIINM,IIIM,IV/V
Wave 0 1998	1050	cigwday	Number cigarettes smoke on weekday
Wave 0 1998	1051	cigwend	Number cigarettes smoke on weekend day
Wave 0 1998	1052	cigdyal	(D) Number of cigarettes smoke a day - inc. non-smokers
Wave 0 1998	1066	smkevr	Whether ever smoked cigarette/cigar/pipe
Wave 0 1998	1067	cignow	Whether smoke cigarettes nowadays
Wave 0 1998	1068	cigevr	Whether ever smoked cigarettes
Wave 0 1998	1069	cigreg	How frequently used to smoke
Wave 0 1998	1070	cigst1	(D) Cigarette Smoking Status - Never/Ex-reg/Ex-occ/Current
Wave 0 1998	1071	cigst2	(D) Cigarette Smoking Status - Banded current smokers

Wave	Variable position	Variable name	Variable definition
Wave 0 1998	1106	ethnicr	HSfE ethnic group collapsed into White and
Wave 0 1998	39	ager	Age last birthday collapsed at 90 plus
Wave 0 1998	41	dobyear	Year of birth collapsed at 90 plus
Wave 0 1998	91	schoh	(D) Social Class of HOH - Harmonised
Wave 0 1998	92	schohg4	(D) Social Class of HOH: I/II,IIINM,IIIM,IV/V
Wave 0 1998	93	schohg6	(D) Social Class of HOH - I,II,IIIN,IIIM,IV,V
Wave 0 1998	94	schohg7	(D) Social Class of HOH - I,II,IIIN,IIIM,IV,V,Others
Wave 0 1999	632	drating	(D) Total Units of alcohol/week
Wave 0 1999	636	overlim	(D) Drinking in relation to weekly limits
Wave 0 1999	649	dnoft	Frequency drank any alcoholic drink last 12 mths
Wave 0 1999	650	dnoft2	(D) Frequency drink alcohol in past 12 months: including non-drinkers
Wave 0 1999	712	dnnow	Whether drink nowadays
Wave 0 1999	713	dnany	Whether drinks occasionally or never drinks
Wave 0 1999	714	dnevr	Whether always non-drinker
Wave 0 1999	78	topqual2	(D) Highest Educational Qualification - Students separate
Wave 0 1999	79	topqual3	(D) Highest Educational Qualification
Wave 0 1999	35	sex	Sex
Wave 0 1999	280	bmival	(D) Valid BMI - inc estimated>130kg
Wave 0 1999	281	bmi	(D) BMI - inc unreliable measurements
Wave 0 1999	282	bmivg4	(D) Valid BMI (grouped:<20,20-25,25-30,30+)
Wave 0 1999	283	bmivg6	(D) Valid BMI (grouped:<20,20-25,25-30,30-35,35-40,40+)
Wave 0 1999	1138	cigdyal	(D) Number of cigarettes smoke a day - inc. non-smokers
Wave 0 1999	1139	cigwday	Number cigarettes smoke on weekday
Wave 0 1999	1140	cigwend	Number cigarettes smoke on weekend day
Wave 0 1999	1145	numsmok	How many cigarettes used to smoke
Wave 0 1999	1153	cigst1	(D) Cigarette Smoking Status - Never/Ex-reg/Ex-occ/Current
Wave 0 1999	1154	cigst2	(D) Cigarette Smoking Status - Banded current smokers
Wave 0 1999	1155	cigsta3	(D) Cigarette Smoking Status: Current/Ex-Reg/Never-Reg
Wave 0 1999	1160	smkevr	Whether ever smoked cigarette/cigar/pipe
Wave 0 1999	1161	cignow	Whether smoke cigarettes nowadays
Wave 0 1999	1162	cigevr	Whether ever smoked cigarettes
Wave 0 1999	1163	cigreg	How frequently used to smoke
Wave 0 1999	1198	ethnicr	HSfE ethnic group collapsed into White and Non-white to avoid disclosure
Wave 0 1999	36	ager	Age last birthday collapsed at 90 plus
Wave 0 1999	39	dobyear	Year of birth collapsed at 90 plus
Wave 0 1999			
Wave 0 2001	1421	drating	(D) Total Units of alcohol/week
Wave 0 2001	1425	overlim	(D) Drinking in relation to weekly limits
Wave 0 2001	1438	dnoft	Frequency drank any alcoholic drink last 12 mths
Wave 0 2001	1439	dnoft2	(D) Frequency drink alcohol in past 12 months: including non-drinkers
Wave 0 2001	1510	dnnow	Whether drink nowadays
Wave 0 2001	1511	dnany	Whether drinks occasionally or never drinks

Wave	Variable position	Variable name	Variable definition
Wave 0 2001	1512	dnevr	Whether always non-drinker
Wave 0 2001	87	topqual2	(D) Highest Educational Qualification - Students separate
Wave 0 2001	88	topqual3	(D) Highest Educational Qualification
Wave 0 2001	64	sex	Sex
Wave 0 2001	1295	bmi	(D) BMI - inc unreliable measurements
Wave 0 2001	1271	nssec8	(D) NS-SEC 8 variable classification (individual)
Wave 0 2001	1272	nssec5	(D) NS-SEC 5 variable classification (individual)
Wave 0 2001	790	smkevr	Whether ever smoked cigarette/cigar/pipe
Wave 0 2001	791	cignow	Whether smoke cigarettes nowadays
Wave 0 2001	792	cigevr	Whether ever smoked cigarettes
Wave 0 2001	794	cigst1	(D) Cigarette Smoking Status - Never/Ex-reg/Ex-occ/Current
Wave 0 2001	795	cigst2	(D) Cigarette Smoking Status - Banded current smokers
Wave 0 2001	796	cigsta3	(D) Cigarette Smoking Status: Current/Ex-Reg/Never-Reg
Wave 0 2001	805	cigwday	Number cigarettes smoke on weekday
Wave 0 2001	806	cigwend	Number cigarettes smoke on weekend day
Wave 0 2001	816	smokyrs	No. of years smoked
Wave 0 2001	1515	ethnicr	HSfE ethnic group collapsed into White and Non-white to avoid disclosure
Wave 0 2001	65	ager	Age last birthday collapsed at 90 plus
Wave 0 2001	67	dobyear	Year of birth collapsed at 90 plus
Wave 1	413	heala	In the past 12 months have you taken an alcoholic drink …? frequency.
Wave 1	4262	edqual	(D) Highest Educational Qualification at ELSA W1
Wave 1	4411	indwei	Sex - Priority: DiSex, DhSex
Wave 1	4414	anssec	FROM HSfE: NS-SEC - long version
Wave 1	4416	enssec	ELSA NS-SEC
Wave 1	404	hesmk	Have you ever smoked cigarettes?
Wave 1	405	heska	Do you smoke cigarettes at all nowadays?
Wave 1	407	heskb	About how many cigarettes a day do you usually smoke on weekdays? instruction to enter midpoint of range given & can~t estimate
Wave 1	410	heskc	About how many cigarettes a day do you usually smoke on weekends? instruction to enter midpoint of range given & can~t estimate
Wave 1	4240	fqethnr	ELSA ethnic group collapsed into White and Non-white to avoid disclosure
Wave 1	4412	indobyr	Year of birth combined HH grid and individual demographics collapsed at 90 plus
Wave 1	4413	indager	Age variable combined info from HH grid and individual demographics collapsed at 90 plus
Wave 2	6536	scako	How often respondent has had an alcoholic drink during the last 12 months
Wave 2	6537	scal7a	Whether respondent had an alcoholic drink in the seven days ending yesterday
Wave 2	6386	FqAQua	Whether has any qualitfications
Wave 2	6387	FqQual1	Further qualifications obtained since last interview (1st mention)

Wave	Variable position	Variable name	Variable definition
Wave 2	6388	FqQual2	Further qualifications obtained since last interview (2nd mention)
Wave 2	6389	FqQual3	Further qualifications obtained since last interview (3rd mention)
Wave 2	6390	fqquzm1	Further qualifications obtained since last interview (1st mention) (merged var)
Wave 2	6391	fqquzm2	Further qualifications obtained since last interview (2nd mention) (merged var)
Wave 2	6568	indsex	Definitive sex variable. Priority: Disex, Dhsex
Wave 2	17	DhSex	Respondent sex from household grid
Wave 2	294	DiSex	Respondent sex
Wave 2	6579	anssec	FROM HSfE: NS-SEC - long version
Wave 2	6581	bnssec	From Wave 1: NS-SEC
Wave 2	731	HeSmk	Whether ever smoked cigarettes
Wave 2	732	HESka	Whether smokes cigarettes at all nowadays
Wave 2	740	HeSkb	Number of cigarettes smoke per weekday
Wave 2	743	HeSkc	Number of cigarettes smoke per weekend day
Wave 2	744	HeTbc	Amount of tobacco smokes per weekend day: whether reported in grams or ounces
Wave 2	6382	fqethnr	Ethnicity recoded into white and non-white
Wave 2	6569	indobyr	Definitive year of birth collapsed at 90 plus. Priority: Didbn, Dhdob
Wave 3	6570	indager	Definitive age variable collapsed at 90 plus.
Wave 3	5888	scako	How often respondent has had an alcoholic drink during the last 12 months
Wave 3	5889	scal7a	Whether respondent had an alcoholic drink in the seven days ending yesterday
Wave 3	6043	w3edqual	(D) Highest Educational Qualification at ELSA Wave 3
Wave 3	290	disex	Respondent sex
Wave 3	5974	indsex	Definitive sex variable. Priority: Disex, Dhsex
Wave 3	26	dhsex	Respondent sex from household grid
Wave 3	6036	w3nssec8	(D) NS-SEC 8 category classification (individual)
Wave 3	6037	w3nssec5	(D) NS-SEC 5 category classification (individual)
Wave 3	831	hesmk	Whether ever smoked cigarettes
Wave 3	832	heska	Whether smokes cigarettes at all nowadays
Wave 3	839	heskb	Number of cigarettes smoke per weekday
Wave 3	842	heskc	Number of cigarettes smoke per weekend day
Wave 3	29	dhager	Age collapsed at 90 plus (use INDAGER instead)
Wave 3	286	diagr	Age from individual demographics collapsed at 90 plus (use INDAGER instead)
Wave 3	5976	indager	Definitive age variable collapsed at 90 plus. Priority: Diag, Dhage
Wave 3	5975	indobyr	Definitive year of birth collapsed at 90 plus. Priority: Didbn, Dhdob
Wave 3	28	dhdobyr	Year of birth collapsed at 90 plus
Wave 4	7838	scako	How often respondent has had an alcoholic drink during the last 12 months
Wave 4	7839	scal7a	Whether respondent had an alcoholic drink in the seven

Wave	Variable position	Variable name	Variable definition
			days ending yesterday
Wave 4	7944	w4edqual	(D) Highest Educational Qualification at ELSA W4
Wave 4	7854	indsex	Definitive sex variable. Priority: Disex, Dhsex
Wave 4	7937	w4nssec5	(D) FINAL W4 NS-SEC 5 category classification (individual)
Wave 4	1076	hesmk	Whether ever smoked cigarettes
Wave 4	1077	heska	Whether smokes cigarettes at all nowadays
Wave 4	7571	fqethnr	Ethnicity recoded into white and non-white
Wave 4	7856	indager	Definitive age variable collapsed at 90 plus. Priority: Diag, Dhage
Wave 4	7855	indobyr	Definitive year of birth collapsed at 90 plus. Priority: Didbn, Dhdob
Wave 4			
Wave 4	Wave 5		
Wave 4	5612	scako	How often respondent has had an alcoholic drink during the last 12 months
Wave 4	5614	scal7b	How many days out of the last seven the respondent had an alcoholic drink
Wave 4	5732	w5edqual	(D) Highest Educational Qualification at ELSA W5
Wave 4	5739	indsex	Definitive sex variable
Wave 4	5725	w5nssec5	(D) FINAL w5 NS-SEC 5 category classification (individual)
Wave 4	1090	hesmk	Whether ever smoked cigarettes
Wave 4	1091	heska	Whether smokes cigarettes at all nowadays
Wave 4	1098	heskb	Number of cigarettes smoke per weekday
Wave 4	5377	fffqethn	Ethnic group (from feed forward information)
Wave 4	5301	fqethnr	Ethnicity recoded into white and non-white
Wave 4	5740	indobyr	Definitive year of birth collapsed at 90 plus. Priority: Didbn, Dhdob
Wave 4	5741	indager	Definitive age variable collapsed at 90+ to avoid disclosure

For cases variables were taken from the Wave that RA diagnosis was firstly reported. For controls the variables were taken from the last Wave (if the variable was missing in that Wave, previous Waves were checked until the value was obtained to maximise the dataset).

Table 81 Final list of selected ELSA variables and their naming

Variable	Wave 0	All other Waves	Final variable name	Missing in all dataset	Missing from respondents with recorded diagnosis of RA
Education	topqual3	edqual	educ	9.60%	1%
Gender	sex	indsex	gender	0%	0%
BMI	bmival*	bmival*	bmi	24.37%	17.17%
Occupational class	schoh**	nssec8**	occup	11.24%	2.79%
Smoking					

Variable	Wave 0	All other Waves	Final variable name	Missing in all dataset	Missing from respondents with recorded diagnosis of RA
Have you ever smoked cigarettes?	cigevr	hesmk	smoke (labels: 0 – never, 1 – ex-smoker)	16.20%	8.22%
Do you smoke cigarettes nowadays?	cignow	heska	smoke (labels: 2 – current smoker)	16.20%	8.22%
Number of cigarettes smoked per week day	cigwday	heskb	smokenum***	Excluded too much missing data	Excluded too much missing data
Number of cigarettes smoked per weekend	cigwend	heskc	Smokenum***	Excluded too much missing data	Excluded too much missing data
Date of birth	dobyear	indobyr	dob	0.67%	0.67%
Age	ager	indager	age	8.70%	0%
Ethnicity	ethnicr	fqethnr	ethn	0.25%	0.20%

6.1.3 Preparing/cleaning ELSA data

Cleaning Wave 0

Wave 0 has information from three different years – 1998, 1999 and 2001. Some variables have different coding at different years (for example occupational class was coded schoh* in 1998 and 1999 and changed to nssec* since 2001). Therefore, variables and their labels were checked and unified as shown in the last column of **Table 81.** Wave 0 from three different years was merged on the principle that a variable was given a value recorded at Wave 0 (1998), if there was a missing value for that variable for the specific person but it was present at Wave 0 (1999), then that value would be assigned; if that value was missing too – Wave 0 2001 value would be assigned. Therefore, every final variable from Wave 0 would start with 'w0variable', indicating that this value is from Wave 0. The data was cleaned - labels (-9,-8, -6, -2 and -1) were changed to missing values.

We used two questions related to smoking. The first was w0hesmk "Have you ever smoked cigarettes?" and w0heska "Do you smoke cigarettes nowadays?". Based on the answers to these questions a new variable **w0smoke** was created to capture smoking status:

- 0 – 'Never smoked' if w0hesmk=2 ('No')
- 1 – 'Ex-smoker' if w0hesmk=1 ('Yes') and w0heska=2 ('No')
- 2 – 'Current smoker' if w0heska=1 ('Yes')

This variable had 29.95% missing data.

Cleaning Wave 1

All variables at Wave 1 were given names starting with 'w1variable'. Occupational info was coded in a different variables anssec and enssec. Enssec variable had 84.99% missing data, while anssec was complete, therefore it was used to generate **w1nssec8**.

W1nssec8 was given labels:

- 1 if anssec labels were any of the following 1, 2, 3.1, 3.2, 3.3, 3.4, 4.1, 4.2, 4.3, 4.4
- 2 if anssec labels were any of the following 5, 6
- 3 if anssec labels were any of the following 7.1, 7.2, 7.3, 7.4
- 4 if anssec labels were any of the following 8.1, 8.2, 9.1, 9.2

- 5 if anssec labels were any of the following 10, 11.1, 11.2
- 6 if anssec labels were any of the following 12.1, 12.2, 12.3, 12.4, 12.5, 12.6, 12.7
- 7 if anssec labels were any of the following 13.1, 13.2, 13.3, 13.4, 13.5
- 8 if anssec label was 14
- 99 if anssec labels were any of the following 15, 16, 17

All the labels can be found in the *Appendix.*

Again we used two questions related to smoking. The first was w1hesmk "Have you ever smoked cigarettes?" and w1heska "Do you smoke cigarettes nowadays?". Based on the answers to these questions a new variable **w1smoke** was created to capture smoking status:

- 0 – 'Never smoked' if w1hesmk=2 ('No')
- 1 – 'Ex-smoker' if w1hesmk=1 ('Yes') and w1heska=2 ('No')
- 2 – 'Current smoker' if w1heska=1 ('Yes')

This variable had 1.58% missing data.

Cleaning Wave 2

All variables at Wave 2 were given names starting with 'w2variable'. At Wave 2 questions about education were only asking about further qualifications (FqAQua, FqQual1-3, fqquzm1-2). ELSA supporting documentation explained this: " FqMqua, FqQual In both HSfE 1998 and 2001, respondents were asked about their qualifications – if we have this information about a respondent he/she will only be asked to report any further qualifications they have obtained since the HSfE interview. Any respondents who were not interviewed at HSfE (and those who were interviewed and refused recontact), will be asked about any qualifications they have ever obtained."

W2edqual was given labels:

- 1 if FqQual1-3 labels were any of the following 1, 23, 24
- 2 if FqQual1-3 labels were any of the following 2-22
- 3 if FqQual1-3 label was 25
- 4 if FqQual1-3 label was 26
- 5 if FqQual1-3 label was 27
- 6 if FqQual1-3 labels were any of the following 28, 29, 95

Fqquzm1-2 were not used as there as 98.84% and 99.93% of data was missing for these variables, respectively.

Occupational info was coded in a different variables anssec and bnssec. bnssec variable had 83.77% missing data, while anssec was complete, therefore it was used to generate **w2nssec8**.

W2nssec8 was given labels:

- 1 if anssec labels were any of the following 1, 2, 3.1, 3.2, 3.3, 3.4, 4.1, 4.2, 4.3, 4.4
- 2 if anssec labels were any of the following 5, 6
- 3 if anssec labels were any of the following 7.1, 7.2, 7.3, 7.4
- 4 if anssec labels were any of the following 8.1, 8.2, 9.1, 9.2
- 5 if anssec labels were any of the following 10, 11.1, 11.2
- 6 if anssec labels were any of the following 12.1, 12.2, 12.3, 12.4, 12.5, 12.6, 12.7
- 7 if anssec labels were any of the following 13.1, 13.2, 13.3, 13.4, 13.5
- 8 if anssec label was 14
- 99 if anssec labels were any of the following 15, 16, 17

All the labels can be found in the *Appendix.*

As in previous wave w2heskb (87.90% missing) and w2heskc (87.91% missing) variables have a lot of missing data. Therefore these questions were not used and we used w2hesmk and w2heska questions. Firstly we used w2hesmk "Have you ever smoked cigarettes?" and w2heska "Do you smoke cigarettes

nowadays?". Based on the answers to these questions a new variable **w2smoke** was created to capture smoking status:

- 0 – 'Never smoked' if w2hesmk=2 ('No')
- 1 – 'Ex-smoker' if w2hesmk=1 ('Yes') and w2heska=2 ('No')
- 2 – 'Current smoker' if w2heska=1 ('Yes')

This variable had 0.06% of missing data.

Cleaning Wave 3

All variables at Wave 3 were given names starting with 'w3variable'.
Variables related to smoking were quite incomplete, w3heskb (88.24% missing), w3heskc (88.25% missing). Therefore we used w3hesmk "Have you ever smoked cigarettes?" and w3heska "Do you smoke cigarettes nowadays?". Based on the answers to these questions a new variable **w3smoke** was created to capture smoking status:

- 0 – 'Never smoked' if w3hesmk=2 ('No')
- 1 – 'Ex-smoker' if w3hesmk=1 ('Yes') and w3heska=2 ('No')
- 2 – 'Current smoker' if w3heska=1 ('Yes')

This variable had 0.05% of missing data.

Cleaning Wave 4

All variables at Wave 4 were given names starting with 'w4variable'. As in the previous Waves w4heskb (89.69% missing) and w4heskc (89.69% missing) variables have lots of missing data. Again we used w4hesmk "Have you ever smoked cigarettes?" and w4heska "Do you smoke cigarettes nowadays?". Based on the answers to these questions a new variable **w4smoke** was created to capture smoking status:

- 0 – 'Never smoked' if w4hesmk=2 ('No')
- 1 – 'Ex-smoker' if w4hesmk=1 ('Yes') and w4heska=2 ('No')
- 2 – 'Current smoker' if w4heska=1 ('Yes')

This variable had 2.02% of missing data.

Cleaning Wave 5

All variables at Wave 5 were given names starting with 'w5variable'. As in previous wave w5heskb (90.83% missing) and w5heskc (90.83% missing) variables have lots of missing data. Therefore, we used w5hesmk "Have you ever smoked cigarettes?" and w5heska "Do you smoke cigarettes nowadays?". Based on the answers to these questions a new variable **w5smoke** was created to capture smoking status:

- 0 – 'Never smoked' if w5hesmk=2 ('No')
- 1 – 'Ex-smoker' if w5hesmk=1 ('Yes') and w5heska=2 ('No')
- 2 – 'Current smoker' if w5heska=1 ('Yes')

This variable had 2.98% of missing data.

6.1.4 Risk factors in ELSA

Table 82 shows appropriate variable/s for identified risk factors of RA. The actual related questions in ELSA for each risk factor are presented in the Appendix (highlighted in yellow).

Table 82 Variables in ELSA related to risk factors

Risk Factor/ELSA variable	Wave 0 1998	Wave 0 1999	Wave 0 2001	Wave1	Wave2	Wave3	Wave4	Wave5
Alcohol	dnevr, dnany, dnnow,	dnevr, dnany, dnnow,	dnevr, dnany,	heala	scako scal7a	scako scal7a	scako scal7a	scako, scal7a

Risk Factor/ELSA variable	Wave 0 1998	Wave 0 1999	Wave 0 2001	Wave1	Wave2	Wave3	Wave4	Wave5
	dnoft, overlim, drating	dnoft, dnoft2, overlim, drating	dnnow, dnoft, dnoft2, overlim, drating					
Blood transfusion	NA	NA	NA	NA	NA	NA	NA	NA
Education	topqual2, topqual3	topqual2, topqual3	topqual2 topqual3	edqual	FqAQua FqQual 1-3, fqquzm 1-2	W3edqual	w4edqual	w5edqual
Coffee consumption	NA	NA	NA	NA	NA	NA	NA	NA
Gender	sex	sex	sex	indsex (DiSex, DhSex)	Indsex (DiSex, DhSex) (sex at Wave 2 nurse)	indsex (dhsex, disex)	Indsex (dhsex at Wave 4 nurse)	indsex
Infections	NA	NA	NA	NA	NA	NA	NA	NA
Obesity/ BMI	bmi, bmival, bmivg4, bmivg6	bmi, bmival, bmivg4, bmivg6	bmi, bmival, bmivg4, bmivg6	NA	bmi, bmival, bmiobe at Wave 2 nurse	NA	bmi, bmival, bmiobe at Wave 4 nurse	NA
Occupational class	schoh, schohg7, schohg6, schohg4	schoh, schohg7, schohg6, schohg4	nssec8, nssec5	anssec, enssec	anssec, bnssec	W3nssec8, w3nssec5	W4nssec8, w4nssec5	w5nssec8, w5nssec5
Reproductive history[29]	Excl.	Excl.	Excl.	Excl.	Excl.	Excl.	Excl.	Excl.
Silica exposure	NA	NA	NA	NA	NA	NA	NA	NA
Smoking	smkevr, cignow[30], cigevr[31], cisgt1, cigst2, numsmok, cigwday, cigwend, cigdyal	smkevr, cignow, cigevr, cisgt1, cigst2, numsmok, cigwday, cigwend, cigdyal	smkevr, cignow, cigevr, cisgt1, cigst2, numsmok, cigwday, cigwend, cigdyal	hesmk, heska, heskb, heskc	HeSkc, HeSkb, HESka, HeSmk	hesmk, heska, heskb, heskc	hesmk, heska, heskb, heskc	hesmk, heska, heskb, heskc
Additional								
Ethnicity	ethnicr	ethnicr	ethnicr	fqethnr	fqethnr	fqethnr	fqethnr	fffqethn fqethnr

[29] It is only related to females, therefore, it will be excluded
[30] cignow corresponds to heska question
[31] Cigevr corresponds to hesmk question

Risk Factor/ELSA variable	Wave 0 1998	Wave 0 1999	Wave 0 2001	Wave1	Wave2	Wave3	Wave4	Wave5
Age	ager, dobyear	ager, dobyear	ager, dobyear	indobyr indager (plus others)	indobyr indager (plus others) (dobyear at Wave 2 nurse)	indager, indobyr (dhdobyr, diagr plus others)	indager, indobyr (plus others) (dobyear at Wave 4 nurse)	indager, indobyr (plus others)

6.2 Further ELSA statistical analyses

The Tables below show outputs from the logistic automatic stepwise forward models using different RA definitions.

Table 83 Automatic stepwise forward logistic model results (excluded respondents that reported OA and hip pain)

Variable	Odds Ratio	95% CI	p-value
Gender			
Male	1.00		
Female	1.33	[1.18-1.5]	<0.001
Ethnicity			
White	1.00		
Non-white	1.77	[1.34-2.33]	<0.001
Education			
NVQ4/NVQ5/Degree or equiv	1.00		
Higher ed below degree	1.30	[1.02-1.67]	0.036
NVQ3/GCE A level equiv	1.00		
NVQ2/GCE O level equiv	1.60	[1.28-1.99]	<0.001
NVQ1/CSE other grade equiv	1.84	[1.36-2.48]	<0.001
Foreign/other	1.52	[1.17-1.98]	0.002
No qualification	2.12	[1.76-2.55]	<0.001
BMI			
<18.4 underweight	1.00		
18.5-24 normal weight	0.72	[0.62-0.84]	<0.001
25-29 overweight	1.00		
>30 obese	1.00		
Smoking status			
Never smoked	1.00		
Ex-smoker	2.29	[1.88-2.8]	<0.001
Current smoker	2.36	[1.89-2.96]	<0.001

Table 84 Automatic stepwise forward logistic model results (excluded respondents with hip pain)

Variable	Odds Ratio	95% CI	p-value
Age			
45-64	1.00		
65-74	1.22	[1.06-1.4]	0.006
75+	1.28	[1.09-1.51]	0.002
Gender			
Male	1.00		
Female	1.42	[1.26-1.6]	<0.001
Ethnicity			
White	1.00		
Non-white	1.53	[1.13-2.06]	0.006
Education			
NVQ4/NVQ5/Degree or equiv	1.00		
Higher ed below degree	1.00		
NVQ3/GCE A level equiv	1.00		
NVQ2/GCE O level equiv	1.27	[1.06-1.53]	0.010
NVQ1/CSE other grade equiv	1.37	[1.04-1.82]	0.028
Foreign/other	1.00		
No qualification	1.55	[1.35-1.78]	<0.001
BMI			
<18.4 underweight	1.00		
18.5-24 normal weight	0.75	[0.64-0.87]	<0.001
25-29 overweight	1.00		
>30 obese	1.00		
Smoking status			
Never smoked	1.00		
Ex-smoker	2.65	[2.14-3.29]	<0.001
Current smoker	2.87	[2.26-3.64]	<0.001

Table 85 Automatic stepwise forward logistic model (excluded respondents with hip pain or hip

Variable	Odds Ratio	95% CI	p-value
Gender			
Male	1.00		
Female	1.37	[1.22-1.55]	<0.001
Ethnicity			
White	1.00		
Non-white	1.51	[1.12-2.04]	0.007
Education			
NVQ4/NVQ5/Degree or equiv	1.00		
Higher ed below degree	1.00		
NVQ3/GCE A level equiv	1.00		
NVQ2/GCE O level equiv	1.36	[1.12-1.65]	0.002
NVQ1/CSE other grade equiv	1.49	[1.11-2]	0.007
Foreign/other	1.30	[1.01-1.67]	0.038
No qualification	1.77	[1.52-2.06]	<0.001
BMI			
<18.4 underweight	1.00		
18.5-24 normal weight	0.76	[0.65-0.89]	0.001
25-29 overweight	1.00		
>30 obese	1.00		
Smoking status			
Never smoked	1.00		
Ex-smoker	2.67	[2.15-3.31]	<0.001
Current smoker	2.78	[2.19-3.54]	<0.001

6.3 ROC curves using different RA definitions

ROC curves were obtained using stepwise forward model for four different RA definitions, described in the sections in the main text.

Table 86 ROC curve obtained using stepwise forward model

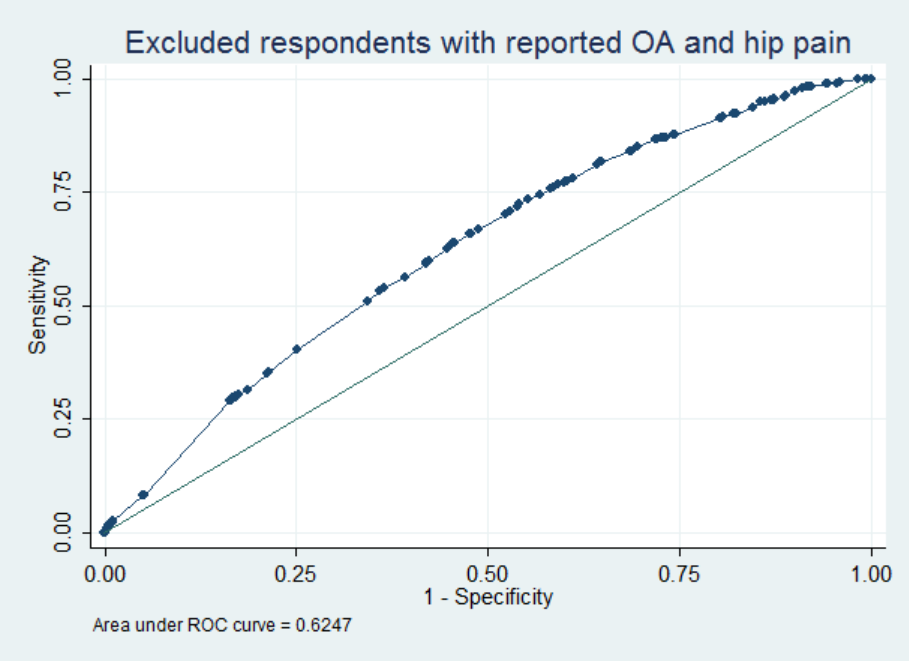


Table 87 ROC curve obtained using stepwise forward model

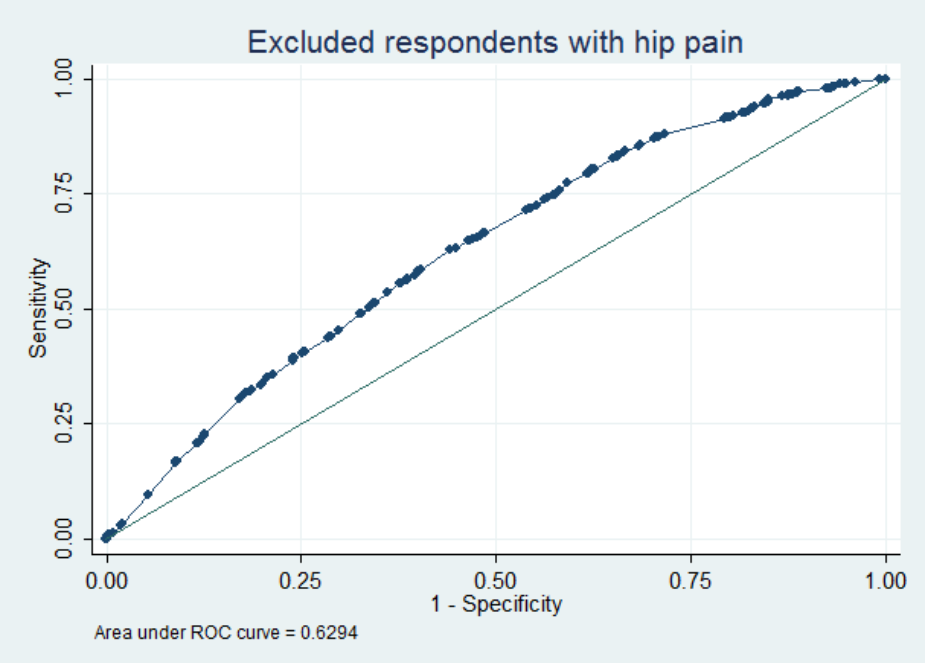
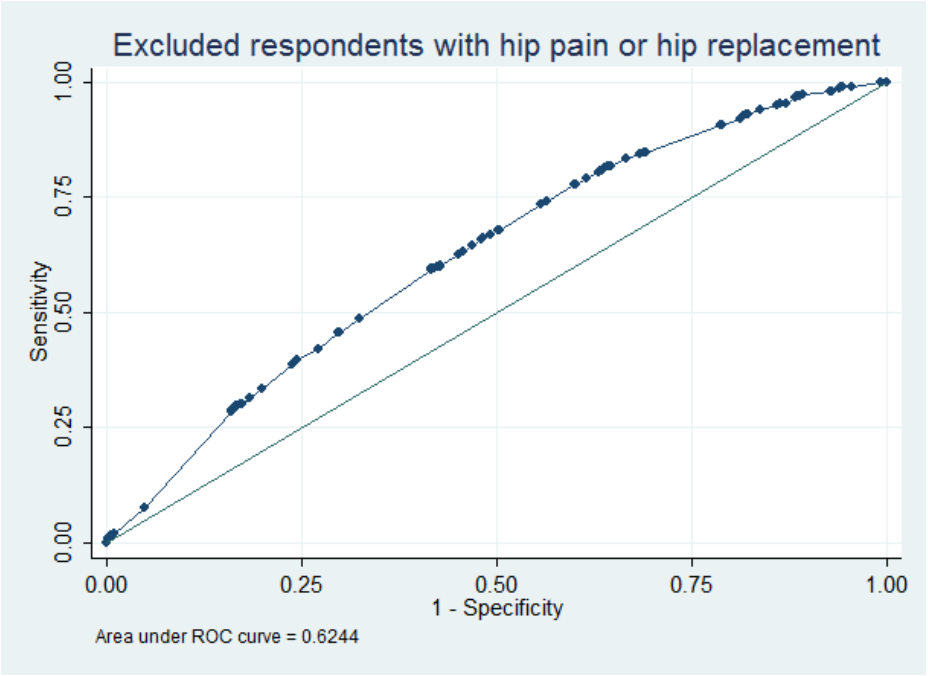


Table 88 ROC curve obtained using stepwise forward model



6.4 Health Survey for England (2005) structure

Figure 30: HSfE 2005 structure

2005 HEALTH SURVEY FOR ENGLAND – CONTENTS

Household data

Household size, composition and relationships	Smoking in household
Accommodation tenure and number of bedrooms	Type of dwelling and area
Economic status/occupation of Household Reference Person	Car ownership
Household income	

Individual level information

	0-1	2-3	4	5-7	8-10	11-12	13-15	16-64	65+ Core	65+ Boost
Interviewer visit										
General health, longstanding illness, limiting longstanding illness, acute sickness, fractures	●	●	●	●	●	●	●	●	●	●
Use of health & dental services									●	●
Use of social care									●	●
Carers responsibilities								●	●	●
CVD, including use of services									●	●
Chronic disease & quality of care									●	●
Disabilities									●	●
Falls									●	●
Physical activity	●	●	●	●	●	●	●			
Smoking					●[a]	●[a]	●[a]	●[b]	●	●[c]
Drinking (seven day period)					●[a]	●[a]	●[a]	●[b]	●	●
Fruit and vegetable consumption				●	●	●	●	●	●	
Eating habits		●	●	●	●	●	●			
Complementary and alternative medicine								●	●	
Economic status/occupation, educational achievement								●	●	●
Ethnic origin	●	●	●	●	●	●	●	●	●	●
Social capital								●[a]	●	●
Height measurement		●	●	●	●	●	●	●	●	●
Weight measurement	●	●	●	●	●	●	●	●	●	●
Reported birth weight	●	●	●	●	●	●	●			
Cycling safety					●[a]	●[a]				
Psychosocial health (GHQ 12)							●[a]	●[a]	●[a]	●[a]
Euroqol general health (EQ5D)								●[a]	●[a]	●[a]
Geriatric depression score									●[a]	●[a]
Social support								●[a]	●[a]	●[a]
Strengths and difficulties			●[e]	●[e]	●[e]	●[e]	●[e]			
Perception of weight					●[a]	●[a]	●[a]			
Use of contraceptive pill								●[a]		
Hormone replacement therapy								●[a,d]		
Incontinence									●[a]	●[a]

[a] These modules were administered by self completion.
[b] This module was administered by self-completion for those aged 16-17 and some aged 18-24.
[c] Shortened smoking module for boost sample 65+
[d] This 18+ only (there are no HRT questions in the young adult self-completion).
[e] This module was asked by proxy and administered by self-completion for parents of 4-15 year olds.

Figure 31 HSfE 2005 nurse visit structure

	0-1	2-3	4	5-7	8-10	11-12	13-15	16-64	65+ Core	65+ Boost
Nurse visit										
Prescribed medicines and vitamin supplements	●	●	●	●	●	●	●	●	●	●
Nicotine replacements								●	●	●
Immunisations	●									
Blood pressure				●	●	●	●	●	●	●
Infant length	●									
Waist and hip circumference						●	●	●	●	●
Demi-span									●	●
Physical function - grip strength, walking speed , balance, chair rise									●	●
Blood sample – total & HDL cholesterol, ferritin, haemoglobin, glycated haemoglobin, fibrinogen, mean corpuscular volume, serum albumin, serum transferrin, vitamin D, vitamin B12									●	●
Saliva sample – cotinine			●	●	●	●	●			
Urine sample								●	●	

6.5 Further HSfE statistical analysis

This section shows ORs for analyses including <44 age group

Table 89 Univariate logistic analysis (including <44 age group)

Variable	Odds Ratio	95% CI	p-value
Age			
<44	1.00		
45-64	2.26	[1.54-3.31]	<0.001
65-74	3.00	[2.03-4.42]	<0.001
75+	2.77	[1.87-4.11]	<0.001
Gender			
Male	1.00		
Female	1.41	[1.28-1.55]	<0.001
Ethnicity			
White	1.00		
Non-white	1.38	[1.12-1.71]	0.003
Education			
NVQ4/NVQ5/Degree or equiv	1.00		
Higher ed below degree	1.84	[1.46-2.31]	<0.001
NVQ3/GCE A level equiv	1.49	[1.14-1.96]	0.004
NVQ2/GCE O level equiv	2.02	[1.63-2.51]	<0.001
NVQ1/CSE other grade equiv	2.24	[1.7-2.94]	<0.001
Foreign/other	2.01	[1.57-2.58]	<0.001
No qualification	2.92	[2.42-3.53]	<0.001
Socioeconomic status			
Higher managerial and professional occup			
Lower managerial and professional occup	1.64	[1.28-2.09]	<0.001
Intermediate occupations	2.35	[1.76-3.15]	<0.001
Small employers and own account workers	2.10	[1.61-2.72]	<0.001
Lower supervisory and technical occup	2.24	[1.75-2.87]	<0.001
Semi-routine occupations	2.72	[2.12-3.48]	<0.001
Routine occupations	2.48	[1.9-3.22]	<0.001
Never worked or long term unemployed	2.39	[1.57-3.64]	<0.001
Other	3.55	[1.62-7.77]	0.002
BMI			
<18.4 underweight			
18.5-24 normal weight	1.65	[0.81-3.39]	0.170
25-29 overweight	2.15	[1.06-4.39]	0.035
>30 obese	2.54	[1.25-5.17]	0.010
Smoking status			
Never smoked			
Ex-smoker	2.68	[2.27-3.17]	<0.001

Current smoker	2.78	[2.31-3.35]	<0.001

Table 90: multivariate logistic regression analysis (including <44 age group)

Variable	Odds Ratio	95% CI	p-value
Age			
<44	1.00		
45-64	2.68	[1.73-4.14]	<0.001
65-74	3.08	[1.97-4.81]	<0.001
75+	2.75	[1.75-4.33]	<0.001
Gender			
Male	1.00		
Female	1.54	[1.38-1.71]	<0.001
Ethnicity			
White	1.00		
Non-white	1.83	[1.43-2.35]	<0.001
Education			
NVQ4/NVQ5/Degree or equiv	1.00		
Higher ed below degree	1.46	[1.12-1.89]	0.005
NVQ3/GCE A level equiv	1.27	[0.94-1.72]	0.123
NVQ2/GCE O level equiv	1.60	[1.25-2.04]	<0.001
NVQ1/CSE other grade equiv	1.75	[1.27-2.4]	0.001
Foreign/other	1.49	[1.12-1.97]	0.006
No qualification	2.03	[1.61-2.56]	0.001
Socioeconomic status			
Higher managerial and professional occup	1.00		
Lower managerial and professional occup	1.26	[0.95-1.66]	0.104
Intermediate occupations	1.50	[1.08-2.09]	0.017
Small employers and own account workers	1.44	[1.07-1.94]	0.017
Lower supervisory and technical occup	1.43	[1.08-1.91]	0.014
Semi-routine occupations	1.55	[1.16-2.06]	0.003
Routine occupations	1.33	[0.98-1.81]	0.068
Never worked or long term unemployed	1.44	[0.89-2.31]	0.136
Other	2.08	[0.83-5.2]	0.117
BMI			
<18.4 underweight	1.00		
18.5-24 normal weight	2.11	[1.03-4.36]	0.043
25-29 overweight	2.91	[1.42-5.98]	0.004
>30 obese	3.34	[1.63-6.84]	0.001
Smoking status			
Never smoked	1.00		
Ex-smoker	2.56	[2.12-3.07]	<0.001
Current smoker	2.72	[2.21-3.34]	<0.001

Table 91: automatic forward stepwise regression analysis (including <44 age group)

Variable	Odds Ratio	95% CI	p-value
Age			
<44	1.00		
45-64	1.00		
65-74	1.16	[1.04-1.31]	0.011
75+	1.00		
Gender			
Male	1.00		
Female	1.52	[1.36-1.69]	<0.001
Ethnicity			
White	1.00		
Non-white	1.81	[1.41-2.31]	<0.001
Education			
NVQ4/NVQ5/Degree or equiv	1.00		
Higher ed below degree	1.59	[1.23-2.05]	<0.001
NVQ3/GCE A level equiv	1.39	[1.03-1.87]	0.030
NVQ2/GCE O level equiv	1.81	[1.43-2.29]	<0.001
NVQ1/CSE other grade equiv	2.03	[1.50-2.75]	<0.001
Foreign/other	1.75	[1.33-2.29]	<0.001
No qualification	2.42	[1.96-2.98]	<0.001
BMI			
<18.4 underweight	1.00		
18.5-24 normal weight	2.05	[0.99-4.22]	0.052
25-29 overweight	2.83	[1.38-5.81]	0.004
>30 obese	3.27	[1.60-6.68]	0.001
Smoking status			
Never smoked	1.00		
Ex-smoker	2.55	[2.12-3.06]	<0.001
Current smoker	2.68	[2.18-3.29]	<0.001

Table 92: automatic backward stepwise regression analysis (including <44 age group)

Variable	Odds Ratio	95% CI	p-value
Age			
<44	1.00		
45-64	2.66	[1.72-4.11]	<0.001
65-74	3.05	[1.95-4.75]	<0.001
75+	2.73	[1.74-4.3]	<0.001
Gender			
Male	1.00		
Female	1.56	[1.4-1.73]	<0.001
Ethnicity			
White	1.00		
Non-white	1.87	[1.46-2.4]	<0.001
Education			
NVQ4/NVQ5/Degree or equiv	1.00		
Higher ed below degree	1.56	[1.21-2.02]	0.001
NVQ3/GCE A level equiv	1.40	[1.04-1.88]	0.027
NVQ2/GCE O level equiv	1.80	[1.42-2.27]	<0.001
NVQ1/CSE other grade equiv	2.00	[1.48-2.72]	<0.001
Foreign/other	1.69	[1.29-2.22]	<0.001
No qualification	2.34	[1.89-2.89]	<0.001
BMI			
<18.4 underweight			
18.5-24 normal weight	2.11	[1.02-4.34]	0.044
25-29 overweight	2.90	[1.42-5.96]	0.004
>30 obese	3.35	[1.63-6.85]	0.001
Smoking status			
Never smoked			
Ex-smoker	2.55	[2.12-3.06]	<0.001
Current smoker	2.73	[2.22-3.36]	<0.001

6.6 CPRD medcodes for joint involvement

Table 93: CPRD medcodes for joint involvement divided into joint groups

Type	Joint	Medcode(s)
Large joints	Shoulder	21524, 24997, 60024, 16166
	Elbow	4228, 17709, 57379, 17001
	Upper arm	29700
	Forearm	51500
	Pelvic region and thigh	68568
	Hip	27394, 53659, 2695
	Knee	443, 17658, 43238, 11569
	Lower leg	34014
	Ankle	25934, 14817, 27746
	Tibio-fibular joint	65998, 107791
	Talonavicular joint	91298
	Sternoclavicular joint	107963
	Acromioclavicular joint	100914
	Sacro-iliac joint	100776
	1st MTP joint	33739
Small joints	Wrist	56187, 48812
	Hand	15570
	MCP joint	48127
	PIP joint of finger	37131
	Distal radio-ulnar joint	94983
	Subtalar joint	94322
	Lesser MTP joint	73723, 99414
	Foot	25934
	IP joint of toe	62465, 107112
Excluded from algorithm	DIP joint of finger	38980
Not site specific		1233, 6892, 6187, 22927, 7404, 1441, 479, 47512, 29396, 3739, 37541, 33506, 1232, 615, 16984, 35448

Table 94: cumulative number of joints involved by type

Type	Joint	Cumulative number of medcodes per patient for each joint									
		1		2		3		4		5+	
		N	%	N	%	N	%	N	%	N	%
Large joints	Shoulder	1001	0.74	189	0.14	39	0.0	12	0.0	5	0.0
	Elbow	946	0.70	209	0.15	23	0.0	14	0.0	6	0.0
	Upper Arm	31	0.02	2	0.00	0	0.0	0	0.0	0	0.0
	Forearm	12	0.01	1	0.00	0	0.0	0	0.0	0	0.0
	Pelvic region	16	0.01	0	0.00	0	0.0	0	0.0	0	0.0
	Hip	304	0.22	35	0.03	7	0.0	3	0.0	0	0.0

Type	Joint	Cumulative number of medcodes per patient for each joint									
		1		2		3		4		5+	
		N	%	N	%	N	%	N	%	N	%
	Knee	2249	16.5	5623	4.13	152	1.1	696	0.5	714	0.5
	Lower Leg	1080	0.79	113	0.08	28	0.0	9	0.0	12	0.0
	Ankle	2043	1.50	441	0.32	58	0.0	29	0.0	13	0.0
	Tibio-fibular	2	0.00	0	0.00	0	0.0	0	0.0	0	0.0
	Talonavicular	4	0.00	0	0.00	0	0.0	0	0.0	0	0.0
	Sternoclavicular	1	0.00	0	0.00	0	0.0	0	0.0	0	0.0
	Acromioclavicul	4	0.00	0	0.00	0	0.0	0	0.0	0	0.0
	Sacro-iliac joint	2	0.00	0	0.00	0	0.0	0	0.0	0	0.0
	1st MTP joint	10	0.01	2	0.00	1	0.0	0	0.0	0	0.0
Small joints	Wrist	23	0.02	4	0.00	0	0.0	0	0.0	0	0.0
	Hand	604	0.44	229	0.17	27	0.0	12	0.0	2	0.0
	MCP joint	12	0.01	2	0.00	0	0.0	0	0.0	0	0.0
	PIP joint of	11	0.01	2	0.00	0	0.0	0	0.0	0	0.0
	Distal radio-	0	0.00	0	0.00	0	0.0	0	0.0	0	0.0
	Subtalar joint	4	0.00	0	0.00	0	0.0	0	0.0	0	0.0
	Lesser MTP	5	0.00	0	0.00	0	0.0	0	0.0	0	0.0
	Foot	53	0.04	3	0.00	1	0.0	0	0.0	0	0.0
	IP joint of toe	11	0.01	0	0.00	0	0.0	0	0.0	0	0.0
Exclude	DIP joint of	7	0.01	0	0.00	0	0.0	0	0.0	0	0.0
No site specified		8137	59.8	1430	10.5	378	2.7	147	1.0	128	0.9

Table 95: number of large joints involved

Number of joints	Frequency	Percentage
0	98,569	72.46
1	27,511	20.22
2	6,684	4.91
3	1,726	1.27
4	783	0.58
5	327	0.24
6	176	0.13
7	94	0.07
8	63	0.05
9	42	0.03
10+	61	0.04
Total	136,036	100.00

6.7 Number of small joints involved

Number of joints	Frequency	Percentage
0	135,034	99.26

Number of joints	Frequency	Percentage
1	717	0.53
2	243	0.18
3	28	0.02
4	12	0.01
5	1	0.00
9	1	0.00
Total	136,036	100.00

Table 96: cross tabulation of number of large and small joints involved

Number of large joints	Number of small joints						Total
	0	1	2	3	4	5+	
0	97,666	634	231	25	11	2	98,569
1	27,428	74	6	2	1	0	27,511
2	6,675	4	5	0	0	0	6,684
3	1,723	2	0	1	0	0	1,726
4	782	1	0	0	0	0	783
5	326	1	0	0	0	0	327
6	175	1	0	0	0	0	176
7	93	0	1	0	0	0	94
8	63	0	0	0	0	0	63
9	42	0	0	0	0	0	42
10+	61	0	0	0	0	0	61
Total	135,034	717	243	28	12	2	136,036

Table 97: joint involvement scores for RA algorithm

Score for RA diagnostic total	Frequency	Percent
0	125,094	91.96
1	9,940	7.31
2	988	0.73
3	14	0.01
Total	136,036	100.00