BMJ open

Representativeness and optimal use of body mass index (BMI) in the UK Clinical Practice Research Datalink (CPRD)

Krishnan Bhaskaran, Harriet J Forbes, Ian Douglas, David A Leon, Liam Smeeth

To cite: Bhaskaran K, Forbes HJ, Douglas I, *et al*. Representativeness and optimal use of body mass index (BMI) in the UK Clinical Practice Research Datalink (CPRD). *BMJ Open* 2013;**3**: e003389. doi:10.1136/bmjopen-2013-003389

▸ Prepublication history and additional material for this paper is available online. To view these files please visit the journal online (http://dx.doi.org/10.1136/bmjopen-2013-003389).

Received 11 June 2013
Revised 6 August 2013
Accepted 12 August 2013

Faculty of Epidemiology and Population Health, London School of Hygiene and Tropical Medicine, London, UK

Correspondence to
Dr Krishnan Bhaskaran; krishnan.bhaskaran@lshtm.ac.uk

ABSTRACT

Objectives: To assess the completeness and representativeness of body mass index (BMI) data in the Clinical Practice Research Datalink (CPRD), and determine an optimal strategy for their use.

Design: Descriptive study.

Setting: Electronic healthcare records from primary care.

Participants: A million patient random sample from the UK CPRD primary care database, aged ≥16 years.

Primary and secondary outcome measures: BMI completeness in CPRD was evaluated by age, sex and calendar period. CPRD-based summary BMI statistics for each calendar year (2003–2010) were age-standardised and sex-standardised and compared with equivalent statistics from the Health Survey for England (HSE).

Results: BMI completeness increased over calendar time from 37% in 1990–1994 to 77% in 2005–2011, was higher among females and increased with age. When BMI at specific time points was assigned based on the most recent record, calendar–year-specific mean BMI statistics underestimated equivalent HSE statistics by 0.75–1.1 kg/m^2. Restriction to those with a recent (≤3 years) BMI resulted in mean BMI estimates closer to HSE (≤0.28 kg/m^2 underestimation), but excluded up to 47% of patients. An alternative strategy of imputing up-to-date BMI based on modelled changes in BMI over time since the last available record also led to mean BMI estimates that were close to HSE (≤0.37 kg/m^2 underestimation).

Conclusions: Completeness of BMI in CPRD increased over time and varied by age and sex. At a given point in time, a large proportion of the most recent BMIs are unlikely to reflect current BMI; consequent BMI misclassification might be reduced by employing model-based imputation of current BMI.

ARTICLE SUMMARY

Strengths and limitations of this study

- The results presented here are based on a large random sample from Clinical Practice Research Datalink (CPRD); therefore, we can confidently generalise the findings to the whole CPRD database and to similar databases based on the UK primary care records.
- To assess the representativeness of CPRD body mass index (BMI) data, we compared with data from the Health Survey for England, which is based on a large nationally representative sample and includes BMI information measured by trained interviewers.
- Our study did not look at BMI recordings among children as this would require a different strategy.

INTRODUCTION

Overweight and obesity are major contributors to global disease burden[1] and are associated with substantial excess mortality.[2] The prevalence of obesity is increasing in developed and developing countries[3 4] and is a growing concern for policy makers. In England, the prevalence of obesity rose steadily from 1993 to 2010: from 13 to 26% in men, and from 16 to 26% in women.[5] Owing to its association with various diseases and health outcomes, body mass index (BMI, the metric most widely used to classify overweight and obesity) is an important factor in many epidemiological studies, both as an exposure and as a potential confounder.

Databases of routinely collected electronic healthcare records are becoming an increasingly valuable resource in epidemiology, allowing population-level research on large, representative samples. The UK Clinical Practice Research Datalink (CPRD) (formerly the General Practice Research Database or GPRD) is widely used and contains comprehensive medical records for approximately 8% of the UK population,[6] allowing epidemiological studies to be carried out on a range of topics and with much greater statistical power than is typically available in traditional cohort studies.

However, a shortcoming of these databases is that lifestyle data, such as BMI, tend to be opportunistically recorded (ie, recorded when the patient is attending for other reasons or when the matter is of direct clinical importance) and can be incomplete. Furthermore, those with non-missing lifestyle data may be unrepresentative of the general population. BMI has been an important covariate in many published studies based on CPRD,[7–14] but the completeness and representativeness of the BMI data have not been previously documented.

Our aim was to undertake an in-depth investigation of BMI recordings in CPRD, including quantifying the completeness of BMI data, and assessing their representativeness by comparing summary statistics based on CPRD data with equivalent statistics from a representative general population survey. We also aimed to suggest and discuss how to deal with the limitations of these routinely collected BMI data.

METHODS

Data sources

Clinical Practice Research Datalink (CPRD

CPRD is a clinical database comprising anonymised computerised medical records from general practitioners (GPs) in the UK. Approximately 8% of the UK population are currently included and the database is broadly representative of the UK population.[15 16] Registration with a GP is near universal in the UK,[17] and GPs act as gatekeepers to the health system so that the CPRD data form a comprehensive health record, comprising demographic information, clinically relevant lifestyle data, prescription details, clinical events, preventive care provided, specialist referrals and hospital admissions and their major outcomes. Data undergo quality checks and practices are designated as 'up to standard' in CPRD from the date that they meet specified data entry quality criteria. For this study, we obtained a random sample of one million CPRD patients, because carrying out the analysis on the full CPRD database would be computationally difficult and the reduction in precision of our estimates that would arise by restricting our analysis to a one million random sample is extremely small.

BMI index data in CPRD

Height and weight measurements are recorded in CPRD whenever measured as part of routine care. We obtained all height and weight records and calculated BMI (BMI=weight/height^2). Records without any measurements or with implausible measurements were excluded (figure 1).

Health survey for England

We obtained published Health Survey for England (HSE) data for BMI from the National Health Service (NHS) Information Centre.[18] HSE is an annual survey designed to produce a representative sample of the adult population aged ≥16 years and living in private households (sample size 14 836 in 2003 and 8420 in 2010). Surveys were interviewer administered with interviewers measuring the weight and height of all participants. Data from 2003 to 2010 were obtained, and these data have been weighted to reduce bias from non-response, based on a logistic regression model incorporating age, sex, household type (based on the number of adults and children living in a household), Strategic Health Authority region and social class (defined using the National Statistics Socioeconomic Classification system). The methods are described in more detail elsewhere.[19]

Statistical methods

Completeness of BMI data in CPRD

In the main analyses, BMI completeness data in CPRD were estimated by calendar period (1990–1994, 1995–1999, 2000–2004, 2005–2011). To calculate completeness for a particular calendar period, all individuals from the one million sample who were registered, aged ≥16 years and under follow-up in 'up to standard' practices on the midpoint of the period were identified and included in the denominator. Among these individuals, the numerator comprised either those with any previous BMI available in their electronic record regardless of how long ago it was entered or those with a BMI available up to 3 years prior to this date. Completeness data were generated by age group, sex and among those for whom, for clinical reasons, BMI should be routinely monitored (those with type 2 diabetes, schizophrenia/other psychoses and ≥2 recent (last 6 months) statin prescriptions). We also investigated whether completeness could be improved by searching for clinical codes ('Read codes') indicating the BMI category. We have not presented CIs for these descriptive statistics because the sample size made sampling error negligible (eg,SEs for the proportions with complete BMI data in the age and calendar-year subgroups were all <0.5%).

Comparison of CPRD BMI data with HSE data

We compared mean BMI over calendar time based on complete CPRD BMI data with equivalent HSE figures, for the period 2003–2010 (since, from 2003 onwards, HSE data were adjusted for non-response). CPRD mean BMI was based on patients registered and under up-to-standard follow-up at the midpoint of the calendar-year. We produced two sets of CPRD mean BMI statistics: first, we used the last BMI observation carried forward (regardless of how long ago it was recorded); second, we restricted our study to patients with a recent BMI available (up to 3 years before the midpoint of the calendar year). As stated above, CIs are not presented because there was negligible sampling error (maximum SE=0.02 kg/m^2). To make like-with-like comparisons with HSE, CPRD data were restricted to English practices (for comparisons with HSE data only), and mean BMI was age-standardised and sex-standardised to the HSE

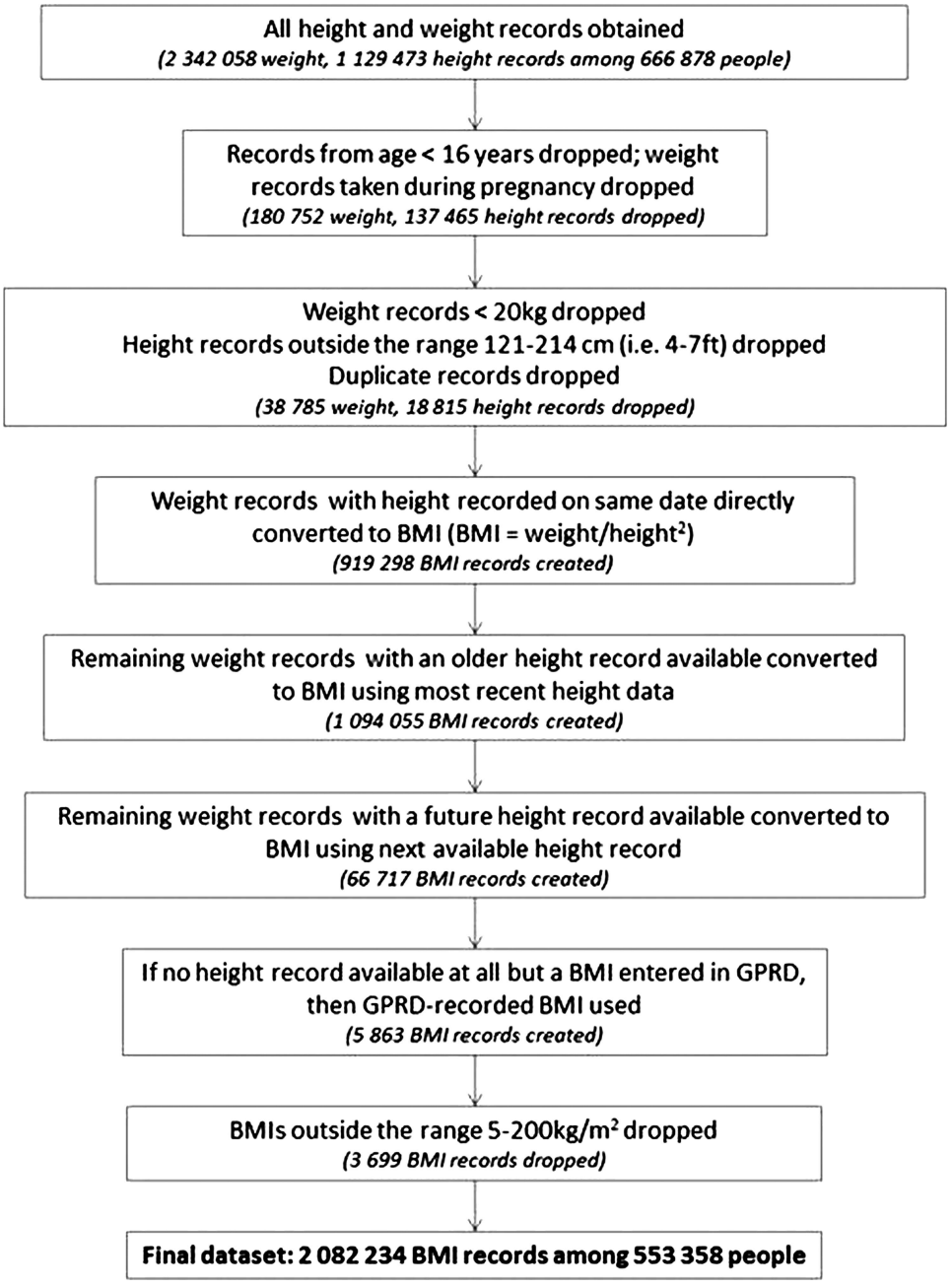


Figure 1 Initial data processing to generate body mass index for analysis.

population structure. Proportions classified as obese (BMI≥30 kg/m^2) over time and based on CPRD and HSE data were also compared.

Model-based imputation of up-to-date BMI measures in CPRD

We explored whether outdated BMI measures in CPRD could be usefully updated by imputation based on a model predicting changes in individual-level BMI over time. We used data from individuals with multiple BMI records to model the expected change in BMI as a function of time since the BMI recording (restricting to individuals with BMI records ≤10 years apart). We fitted a linear regression model with change in BMI as the outcome; the main covariate predicting change in BMI was elapsed time, which was included as a three knot cubic spline to allow for non-linearity; we also included interactions between the spline basis variables and indicator variables for age and sex. Feasible weighted least squares estimation was used to allow for heteroscedasticity.[20]

Having specified a model for change in BMI over time, we first explored its performance among individuals with at least two BMIs entered in CPRD, by predicting the most recent BMI based on the previous BMI record and the elapsed time; we compared the distribution of the errors from this approach with the distribution of the errors by simply using the last observation carried forward. We then repeated the comparison with the HSE mean BMI data for each calendar year. This time, we

Table 1 Completeness of BMI data in the CPRD, by age and calendar period

Age group (years)	1990–1994	1995–1999	2000–2004	2005–2011
16–24*				
N registered	11 423	17 501	34 452	42 546
BMI in the previous 3 years (%)	26	28	25	32
BMI in the previous (%)	26	37	30	37
25–34				
N registered	17 477	29 923	48 659	50 413
BMI in the previous 3 years (%)	37	39	36	49
BMI in the previous (%)	38	66	67	72
35–44				
N registered	15 953	28 838	55 991	61 014
BMI in the previous 3 years (%)	36	36	31	46
BMI in the previous (%)	39	67	71	80
45–54				
N registered	14 507	27 765	48 093	55 564
BMI in the previous 3 years (%)	39	37	32	50
BMI in the previous (%)	42	70	73	84
55–64				
N registered	11 680	20 843	42 258	49 380
BMI in the previous 3 years (%)	42	40	37	57
BMI in the previous (%)	44	74	77	87
65–74				
N registered	10 678	17 605	30 997	34 508
BMI in the previous 3 years (%)	36	37	40	67
BMI in the previous (%)	38	71	79	91
75+				
N registered	8637	16 005	29 384	32 523
BMI in the previous 3 years (%)	28	32	37	64
BMI in the previous (%)	28	56	69	87
Total				
N registered	90 355	158 480	289 834	325 948
BMI in the previous 3 years (%)	35	36	34	51
BMI in the previous (%)	37	64	67	77

N registered is all those under follow-up at the midpoint of the period.
*BMI measurements from age <16 years were not counted in this analysis; hence, completeness in the 16–24 age group may be artificially low.

included all individuals with a BMI record in the previous 10 years and used the model described above to impute current BMI at the midpoint of the calendar year by predicting the change in BMI since the last available BMI record. We did this within a multiple imputation framework (using five imputations) to account for uncertainty in the modelled changes over time.[21]

RESULTS

Completeness of BMI data in CPRD

In 1990–1994, 37% of individuals had at least one previously recorded BMI, and the proportion increased to 77% by 2005–2011(table 1).The proportion of individuals with a recent BMI (recorded in the previous 3 years) was lower in each calendar-period (35% in 1990–1994 rising to 51% in 2005–2011). BMI completeness generally increased with age up to 75 years, with a lower proportion in the oldest age group having data available. Data for single calendar years are shown in online supplementary appendix table A1 and illustrate similar patterns. BMI data appeared to be consistently more widely available among women than men (figure 2). As expected, BMI completeness was higher in particular

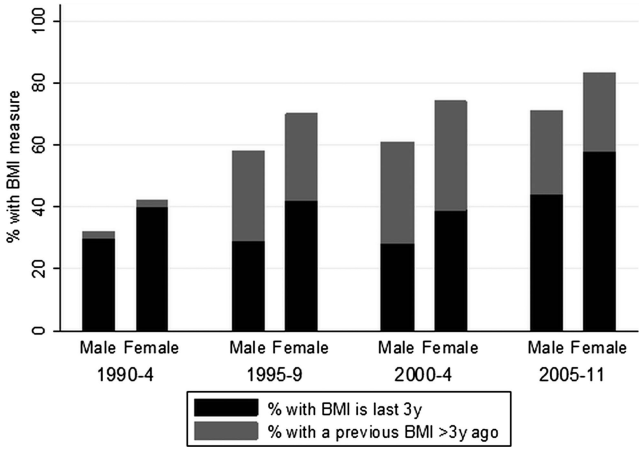


Figure 2 Completeness of body mass index data in Clinical Practice Research Datalink, by gender and calendar–period.

clinical subgroups: in total, 97% of patients with a record of type II diabetes had a recent BMI recorded, along with over 78% of those with a diagnosis of schizophrenia/psychoses (see online supplementary appendix table A2). This is in line with the quality and outcomes framework (QOF), which has encouraged BMI monitoring in these conditions since 2004.[22] BMI completeness was also high among current statin users (82% with a recent BMI available).

There was little extra information available in clinical ('Read') codes relating to BMI. In the most recent calendar period, of 75 518 individuals with no previous BMI record available, only 1222 (1.6%) had ever had a clinical code that would enable classification into BMI categories (underweight, normal, overweight/obese). Furthermore, for those with a previous BMI, only a small proportion had more recent information related to BMI recorded in a clinical code (7675/250 430=3.0% in the most recent period).

Summary statistics using complete CPRD BMI data and comparison with HSE

We found that age-standardised and sex-standardised mean BMI based on CPRD data was consistently and substantially lower (by up to 1.1 kg/m^2) than in the HSE data (mean BMI in CPRD=25.7 kg/m^2 in 2003 rising to 26.3 in 2010, compared with 26.8 kg/m^2 (95% CI 26.7 to 26.9) and 27.3 (27.1 to 27.5), respectively, in HSE; figure 3).

When BMI entries more than 3 years old were discarded, between 33 and 47% of patients were lost across calendar-years. However, the estimated mean BMI in CPRD was considerably closer to what would be expected based on the HSE data, with the CPRD data underestimating the HSE statistics by only between 0.04 and 0.28 kg/m^2 in individual calendar-years and the CPRD estimate falling within the HSE CI for two of the most recent 3 calendar-years (mean BMI in CPRD=26.9, 27.0 and 27.0 kg/m^2 compared with 27.0 (26.9 to 27.1), 27.0 (26.8 to 27.2) and 27.3 (27.1 to 27.5) in HSE, in 2008, 2009 and 2010, respectively). Age-stratified and sex-stratified data demonstrated similar patterns, except that in the eldest age group (75+ years), restriction to those with recent BMI measures did not bring the estimated BMI substantially closer to the HSE figures (see online supplementary appendix figure A1).

We also compared the proportions classified as obese between CPRD and HSE (see online supplementary appendix figure A2). Consistent with the previous analysis, using any previous BMI reading to classify individuals in CPRD resulted in lower obesity rates than expected based on the HSE data, while restricting to patients with a recent reading led to estimated obesity rates close to those in HSE.

Model-based imputation of up-to-date BMI measures in CPRD

The contrast between BMI summary statistics based on recent measures and those based on any previous measures suggested that older BMI records were tending to underestimate current BMI. We therefore examined whether a model could be developed to impute current BMI, taking into account the elapsed time since the last measure. In a linear regression model for change in BMI over time, we estimated that on average BMI increased over the 10-year period following a BMI record for those aged up to 69 years at the time of the record and decreased over time in those aged 70 years or more (see online supplementary appendix figure A3). We tested the predictive performance of our model by predicting the most recent BMI based on the previous one, among patients with CPRD with more than one recorded BMI available. When the older BMI was less than 3 years old, there was little gain in applying the correction compared with carrying the older observation forward (figure 4). However, when there was a longer

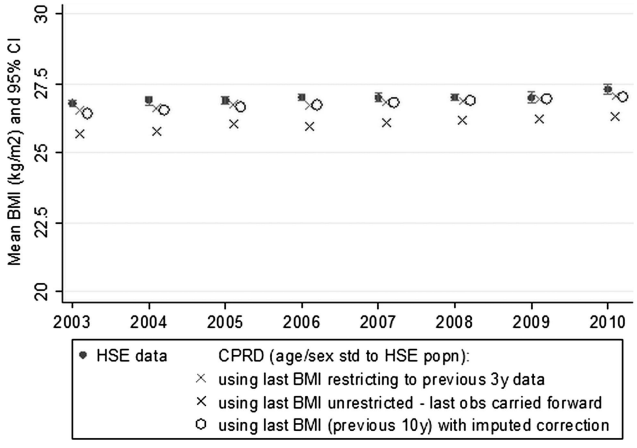


Figure 3 Mean body mass index (BMI) over calendar time comparing those with BMI recorded in Clinical Practice Research Datalink (English practices) with the Health Survey for England 2010 data.

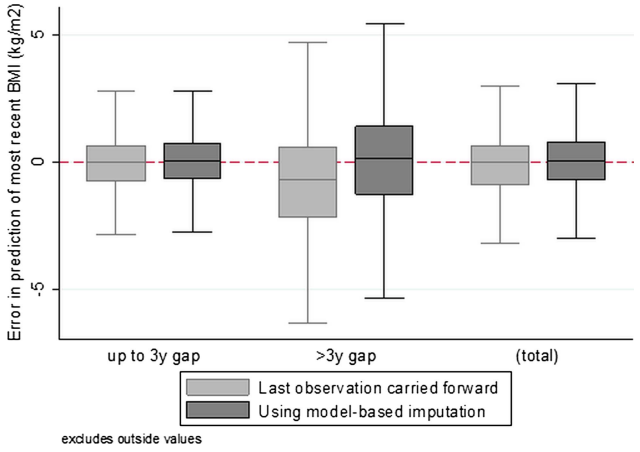


Figure 4 Error in prediction of most recent BMI from older BMI, comparing simple last observation carried forward with model-based imputation of up-to-date BMI—stratified by time gap between readings.

gap, carrying the previous BMI forward tended to underestimate the later BMI, while employing the model-based imputation removed the underestimation and led to smaller errors on average (median error= −0.70 kg/m^2 (IQR −2.18 to +0.56) using the last observation carried forward, compared with +0.11 kg/m^2 (−1.29 to +1.40) using the model-based imputation).

We then repeated the comparison of mean BMI in CPRD versus HSE, this time using our model for change in BMI over time as a basis for performing multiple imputations of current BMI based on the latest available measure and the time since it was recorded. The estimated mean BMIs were now in line with those based on only recent data in the earlier analysis, being only between 0.04 and 0.37 kg/m^2 lower than the HSE statistics in individual calendar years (figure 3, circles). Even with multiple imputation, CIs remained extremely narrow (<0.07 kg/m^2) due to the large sample size, and therefore are not shown in the figure. Of note, all patients with a BMI recorded up to 10 years before the midpoint of the calendar year of interest were now included in the estimation of the 'corrected' means; thus, in individual calendar years, only 9–13% of patients were dropped, compared to 33–47% of patients when dropping BMI records >3 years old.

DISCUSSION

Main findings

BMI completeness has increased over calendar time (rising from 37% in 1990–1994 to 77% in 2005–2011). Completeness was higher among females, older age groups and clinical subgroups where recording BMI is encouraged. When BMI on the date of interest was assigned to individual patients in CPRD using the last available record, regardless of how long ago it was entered, we found that the resulting mean BMI statistics for the CPRD population were consistently lower than the equivalent HSE estimates (by up to 1.1 kg/m^2). This appeared to be driven by older BMI records tending to systematically underestimate current BMI: when only recent CPRD BMI records (≤3 years old) were used, the mean BMI statistics were closer to the HSE estimates. However, a substantial number of patients were then excluded altogether (33–47% across years). Finally, we suggested a process for modelling changes in BMI after a BMI record, which could allow researchers to impute BMI on the date of interest and avoid dropping large numbers without a recent measure from their analyses.

Comparison with other studies

There are very few comparable studies (see online supplementary appendix table A2). However, the proportion of patients with a recent BMI recording in CPRD is in line with a summary of the QRESEARCH database (a similar UK primary care database with data from over 530 general practices using EMIS software rather than VISION software)[23]; by March 2007, 58% of registered patients aged 16+ years had their BMI recorded in the past 5 years; this compares with 51% with a BMI recorded in the last 3 years in our analysis (for 2005–2011). Similar to our study, the QRESEARCH report showed an increase in completeness over time, rising from 42% in 2000/2001 to 58% in 2007. In a third UK primary care database, The Health Improvement Network (THIN) , the proportion of newly registered patients between 2004 and 2006 with BMI data was in line with our findings; 62% of patients had a height recording and 66% had a weight recording within 12 months of registration.[24]

Explanation of findings

Completeness

Increasing the completeness of BMI over time may reflect a general trend towards encouragement to record BMI in primary care. Greater BMI completeness among females and older age groups may have a number of explanations including higher consultation rates in primary care [25 26] and different prevalences of diseases in which it is important to monitor BMI.

Comparison of CPRD BMI data with HSE data

Mean BMI based on the CPRD population was lower in each calendar year than the equivalent HSE estimates when BMI in CPRD was assigned using the last available record; however, when the analysis was restricted to those with a recent BMI record, estimates from CPRD were close to the HSE estimates. This suggests that the substantial proportion of BMI recordings in CPRD that were outdated on the date of interest may have driven the apparent underestimation of mean BMI in CPRD in the unrestricted analysis. This, in turn, would imply that individual BMIs tend to increase over time, and indeed when we specifically modelled changes in BMI over time, we found a pattern of increasing BMI with age for those <70 years old, consistent with prospective cohort studies with repeated BMI measurements[27–29]; this pattern of increasing BMI over time is likely to be driven specifically by weight change, since adult height would not change substantially in this age range. A simple adjustment of outdated BMIs based on our modelled changes over time brought the CPRD mean BMI statistics in line with the HSE estimates, and when we validated the adjustment in a subset of patients with repeated BMI measures, we found smaller errors on average, compared with simply carrying outdated BMI records forwards.

Of note, we observed that CPRD consistently underestimated BMI compared to HSE among those aged ≥75 years, even when only recent records were used; this may reflect the fact that institutionalised patients are represented in CPRD but not in HSE: HSE may not be an ideal comparison for this age group since elderly people in institutions (who are represented in CPRD) may be more likely to be frail and have lower BMIs than those living in private households.

Implications

First, our findings suggest that BMI completeness is likely to vary between studies depending on the study population and study period. BMI data are not likely to be missing completely at random (eg, missingness may vary by patient characteristics or particular diseases). There may be information in the database, however, which predicts missingness and which could satisfy the 'missing at random' assumption required for multiple imputation. A study exploring the potential of imputing missing data in THIN found that after multiple imputation, summary statistics of height and weight were comparable with data from nationally representative datasets.[24]

Second, our analyses suggest that the common practice of assigning BMI status based on the nearest/most recently available record to the index date of interest might lead to misclassification, given that a large number of patients have only substantially outdated BMI records available at any particular time. Strategies to address this include restricting to recent BMI, but this is likely to exclude a large number of patients. We have suggested an alternative strategy based on updating the outdated BMIs by modelling changes in BMI over time, though this is not without drawbacks: the approach requires an assumption that individuals with ≥2 BMI records available (needed to estimate the model for changes over time) are representative of the wider patient population, which may not be the case; it is also a more complex strategy, particularly if performed within a multiple imputation framework to allow for uncertainty in the correction, which could be substantial in studies with smaller sample sizes than considered here. Other imputation strategies could also be considered in certain contexts, such as the 2-fold algorithm, which imputes missing data from longitudinal variables at particular time points by using adjacent data points.[30] Ultimately, the pros and cons of various methods, as well as the optimal strategy to use, are likely to depend on the particular study and the characteristics of the study population.

Strengths and limitations

The results presented here are based on a large random sample from CPRD; therefore, we can confidently generalise the findings to the whole CPRD database. Although we cannot assume that these findings will relate to UK routinely collected primary care databases based on other IT systems (CPRD is based on practices using VISION), the underlying processes driving BMI recording are likely to be similar. This study did not look at BMI recordings among children as this would require a different strategy. Completeness among the 16–24 years of age group may be artificially low because weights recorded at age <16 were excluded, so those at the lower end of the age group will not have had as much time to accrue weight recordings. We believe HSE to be the best available comparison for this study; it is a nationally representative, large sample utilising height and weight recordings measured by a trained interviewer and is weighted for non-response.[19 31] However, there is a degree of missing data in HSE, which is a limitation. In 2010, just over 85% of adults interviewed provided valid height and weight recordings.[29] One of the most common reasons for missing BMI was refusal (up to 8% were missing due to refusal),[19] which if related to BMI status may bias the estimates of mean BMI in HSE. Our comparisons between CPRD-based and HSE-based BMI statistics focused on the mean (and in the online supplementary appendix, on the proportion classed as obese); these are the principal statistics published in the HSE trend tables, so we were not able to look at a broader range of measures of the BMI distribution that might be of interest to researchers using BMI data in the context of public health. Finally, we have not attempted to quantify or comment on the usefulness of BMI as a measure of adiposity, and researchers using BMI data should consider whether it is the best available measure for their purposes.

CONCLUSIONS

Completeness of BMI data in CPRD varies over time and by age and sex. BMI records may become outdated over time and naive use could lead to misclassification of BMI status. We used a 3-year cut-off to define a recent BMI; further research could include a systematic analysis of how long BMI records can be considered 'up-to-date', and whether this varies by patient characteristics. The optimal strategy for assigning BMI status to individuals in studies based on CPRD and similar electronic healthcare databases is likely to depend on the specific study population and the research context.

Contributors KB developed the analytical strategy for this paper, processed and analysed the data and wrote the paper. HF was involved in discussing the data processing and analysis of the data, as well as the writing of the paper. LS was involved in discussions of the analytical approach to this study and made comments on the analysis and the writing of the paper. ID was involved in discussions of the analytical approach to this study and made comments on the analysis and the writing of the paper. DL was involved in discussions of the analytical approach to this study and made comments on the analysis and the writing of the paper.

Funding This report is independent research arising from a postdoctoral fellowship (for KB) supported by the National Institute for Health Research (PDF-2011-04-007). ID is supported by an MRC methodology research fellowship. LS is supported by a Wellcome Trust senior research fellowship in clinical science.

Competing interests None.

Ethics approval The study was approved by the London School of Hygiene and Tropical Medicine Ethics Committee. MHRA Independent Scientific Advisory Committee.

Provenance and peer review Not commissioned; externally peer reviewed.

Data sharing statement This analysis is based on a large random sample from the Clinical Practice Research Datalink, provided by the UK Medicines and Healthcare products Regulatory Agency. The authors' licence for using these data does not allow sharing of raw data with third parties.

REFERENCES

1. World Health Organisation. *Global health risks: mortality and burden of disease attributable to selected major risks*. Geneva, Switzerland: World Health Organisation, 2009.
2. Flegal KM, Graubard BI, Williamson DF, *et al*. Excess deaths associated with underweight, overweight, and obesity. *JAMA* 2005;293:1861–7.
3. Swinburn BA, Sacks G, Hall KD, *et al*. Obesity 1 The global obesity pandemic: shaped by global drivers and local environments. *Lancet* 2011;378:804–14.
4. Kelly T, Yang W, Chen CS, *et al*. Global burden of obesity in 2005 and projections to 2030. *Int J Obesity* 2008;32:1431–7.
5. NHS Information Centre. Health survey for England—2010: health and lifestyles. Secondary Health Survey for England—2010: health and lifestyles 2011. http://www.ic.nhs.uk/pubs/hse10report
6. CPRD. Clinical Practice Research Database (CPRD) website. Secondary Clinical Practice Research Database (CPRD) website. http://www.cprd.com/intro.asp
7. Delaney JA, Daskalopoulou SS, Brophy JM, *et al*. Lifestyle variables and the risk of myocardial infarction in the general practice research database. *BMC Cardiovasc Disord* 2007;7:38.
8. Green J, Czanner G, Reeves G, *et al*. Oral bisphosphonates and risk of cancer of oesophagus, stomach, and colorectum: case-control analysis within a UK primary care cohort. *BMJ* 2010;341:c4444.
9. Tzoulaki I, Molokhia M, Curcin V, *et al*. Risk of cardiovascular disease and all cause mortality among patients with type 2 diabetes prescribed oral antidiabetes drugs: retrospective cohort study using UK general practice research database. *BMJ* 2009;339:b4731.
10. Douglas I, Smeeth L, Irvine D. The use of antidepressants and the risk of haemorrhagic stroke: a nested case control study. *Br J Clin Pharmacol* 2011;71:116–20..
11. Andersohn F, Schade R, Suissa S, *et al*. Long-term use of antidepressants for depressive disorders and the risk of diabetes mellitus. *Am J Psychiatry* 2009;166:591–8.
12. Lawrenson R, Todd JC, Leydon GM, *et al*. Validation of the diagnosis of venous thromboembolism in general practice database studies. *Br J Clin Pharmacol* 2000;49:591–6.
13. Jick H, Zornberg GL, Jick SS, *et al*. Statins and the risk of dementia. *Lancet* 2000;356:1627–31.
14. van Staa TP, Wegman S, de Vries F, *et al*. Use of statins and risk of fractures. *JAMA* 2001;285:1850–5.
15. Office for National Statistics. *Key Health Statistics from General Practice 1998:* Analyses of morbidity and treatment data, including time trends, England and Wales. London: Office for National Statistics, 2000.
16. Parkinson JP, Davis S, Van Staa T. The General Practice Research Database: now and the future. In: Mann R, Andrews EB, eds. *Pharmacovigilance*. Chichester: John Wiley and Sons, 2007:341–8.
17. Schoonen WM, Thomas SL, Somers EC, *et al*. Do selected drugs increase the risk of lupus? A matched case-control study. *Br J Clin Pharmacol* 2010;70:588–96.
18. NHS Information Centre. Health Survey for England—2010: trend tables. Secondary Health Survey for England—2010: trend tables. http://www.ic.nhs.uk/statistics-and-data-collections/health-and-lifestyles-related-surveys/health-survey-for-england/health-survey-for-england–2010-trend-tables
19. Aresu M, Boodhna G, Bryson A, *et al*. Volume 2: Methods and documentation. In: Craig R, Mindell J, eds. *Health Survey for England 2010*. Leeds: NHS Information Centre for Health and Social Care, 2011.
20. Greene WH. *Econometric analysis*. Upper Saddle River, NJ: Prentice Hall, 1997.
21. Rubin DB. *Multiple imputation for nonresponse in surveys*. New York: J Wiley & Sons, 1987.
22. NHS. *Quality and outcomes framework guidance for GMS contract 2011/12: employers and British Medical Association*. 2011.
23. NHS Information Centre. A summary of public health indicators using electronic data from primary care. Secondary A summary of public health indicators using electronic data from primary care 2008. http://www.ic.nhs.uk/article/2021/Website-Search?productid=4287&q=A+summary+of+public+health+indicators+using+electronic+data+from+primary+care.&sort=Relevance&size=10&page=1&area=both#top
24. Marston L, Carpenter JR, Walters KR, *et al*. Issues in multiple imputation of missing data for large general practice clinical databases. *Pharmacoepidemiol Drug Saf* 2010;19:618–26.
25. Rowlands S, Moser K. Consultation rates from the General Practice Research Database. *Br J Gen Pract* 2002;52:658–60.
26. The Health and Social Care Information Centre. Trends in Consultation Rates in General Practice 1995 to 2009. Secondary Trends in Consultation Rates in General Practice 1995 to 2009. 2009. http://www.ic.nhs.uk/article/2021/Website-Search?productid=729&q=qresearch&sort=Relevance&size=10&page=1&area=both#top
27. Li L, Law C, Power C. Body mass index throughout the life-course and blood pressure in mid-adult life: a birth cohort study. *J Hypertens* 2007;25:1215–23.
28. Silverwood RJ, Pierce M, Thomas C, *et al*. Association between younger age when first overweight and increased risk for CKD. *J Am Soc Nephrol* 2013;24:813–21.
29. Tirosh A, Shai I, Afek A, *et al*. Adolescent BMI trajectory and risk of diabetes versus coronary disease. *N Engl J Med* 2011;364:1315–25.
30. Welch C, Bartlett J, Petersen I. Application of multiple imputation using the two-fold fully conditional specification algorithm in longitudinal clinical data. *Stata J* 2013 (In Press).
31. Mindell J, Biddulph JP, Hirani V, *et al*. Cohort Profile: The Health Survey for England. *Int J Epidemiol* 2012:1–9.

Appendix Table A1: Completeness of BMI data in CPRD, by age and calendar year (men and women combined)

	1990	1991	1992	1993	1994	1995	1996	1997	1998	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011
16-24*																						
N registered	8009	10620	11423	12525	12959	13461	14711	17501	19447	23227	28581	31794	34452	36191	38017	39512	40433	41635	42546	42646	42216	41093
BMI in previous 3y (%)	11	20	26	32	34	35	32	28	25	24	23	24	25	26	27	28	30	32	32	34	33	32
BMI any previous (%)	11	20	26	34	39	40	39	37	34	32	30	30	30	31	32	33	34	36	37	39	39	38
25-34																						
N registered	11065	15482	17477	20071	21596	22694	25023	29923	32560	37448	44253	46730	48659	48853	49901	50695	50271	50586	50413	50809	50435	50192
BMI in previous 3y (%)	15	27	37	44	46	47	43	39	36	35	34	35	36	38	41	42	45	47	49	51	50	50
BMI any previous (%)	15	27	38	49	57	62	64	66	66	67	66	66	67	67	68	69	70	71	72	73	73	73
35-44																						
N registered	10912	14569	15953	18120	19383	20797	23403	28838	32305	38809	47168	51881	55991	58121	60408	61817	61912	61983	61014	59595	57197	54926
BMI in previous 3y (%)	15	26	36	44	45	45	41	36	32	30	29	29	31	33	36	38	41	44	46	48	47	47
BMI any previous (%)	15	27	39	50	58	64	66	67	68	69	69	69	71	72	74	75	77	79	80	81	82	83
45-54																						
N registered	8732	12461	14507	17033	18662	20199	22946	27765	30629	35954	43045	45882	48093	49276	51084	52921	53519	54738	55564	56368	56266	56070
BMI in previous 3y (%)	17	29	39	47	47	47	42	37	33	31	30	30	32	34	38	41	44	48	50	51	50	50
BMI any previous (%)	17	30	42	53	61	66	69	70	71	72	72	72	73	74	76	78	80	82	84	85	86	86
55-64																						
N registered	7613	10263	11680	13329	14310	15289	17178	20843	23615	28115	33854	38058	42258	44405	46749	48398	49162	49346	49380	49207	48077	46430
BMI in previous 3y (%)	17	31	42	51	52	51	45	40	37	35	34	35	37	40	45	48	51	56	57	58	56	57
BMI any previous (%)	17	32	44	57	66	70	72	74	74	75	75	76	77	79	81	83	84	86	87	88	89	90
65-74																						
N registered	6715	9346	10678	12190	13128	13554	14900	17605	19246	22460	26600	28870	30997	31790	32914	33516	33402	33895	34508	35245	34845	35163
BMI in previous 3y (%)	14	26	36	45	47	47	42	37	35	34	33	35	40	45	52	56	61	65	67	67	66	67
BMI any previous (%)	14	27	38	51	60	65	69	71	73	75	76	77	79	81	84	87	88	90	91	92	93	93
75+																						
N registered	5680	7796	8637	9926	10670	11852	13416	16005	17989	20997	25213	27585	29384	30016	31036	31849	32157	32484	32523	32507	32007	31621
BMI in previous 3y (%)	9	20	28	34	35	37	35	32	31	31	31	33	37	41	48	54	58	62	64	67	66	67
BMI any previous (%)	9	20	28	37	43	49	53	56	59	61	63	65	69	72	76	80	83	85	87	89	90	91
Total																						
N registered	58726	80537	90355	103194	110708	117846	131577	158480	175791	207010	248714	270800	289834	298652	310109	318708	320856	324667	325948	326377	321043	315495
BMI in previous 3y (%)	14	26	35	43	45	45	40	36	33	31	31	31	34	36	40	43	46	50	51	52	52	52
BMI any previous (%)	15	27	37	48	56	61	63	64	65	66	65	66	67	69	71	72	74	76	77	78	79	79

Appendix Table A2: Completeness of BMI data among patients with diagnoses of type 2 diabetes or mental health disorders, and among statin users

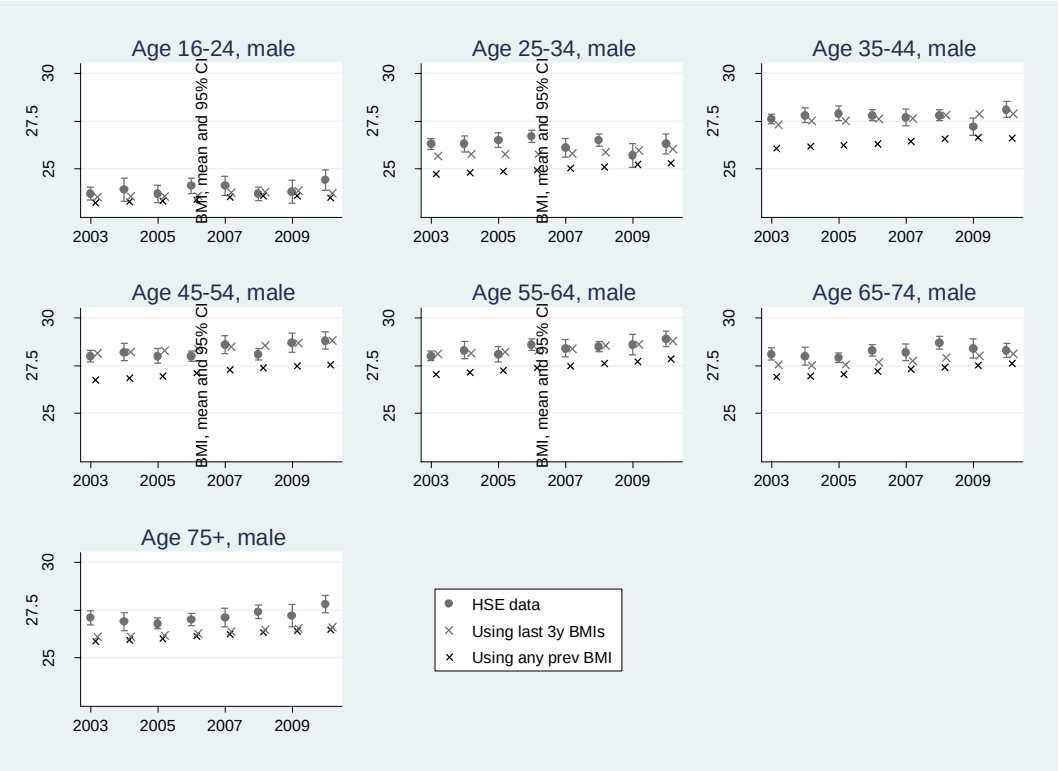
	Type 2 diabetes		Schizophrenia/psychoses		Statin users	
	N registered	N with BMI in last 3y (%)	N registered	N with BMI in last 3y (%)	N registered	N with BMI in last 3y (%)
1990	542	172 (32)	168	32 (19)	18	6 (33)
1991	773	382 (49)	234	62 (26)	63	27 (43)
1992	987	558 (57)	282	108 (38)	137	99 (72)
1993	1239	849 (69)	317	133 (42)	183	142 (78)
1994	1400	1080 (77)	352	156 (44)	233	174 (75)
1995	1636	1311 (80)	388	184 (47)	344	233 (68)
1996	1991	1560 (78)	423	184 (43)	692	408 (59)
1997	2553	1982 (78)	514	181 (35)	1296	688 (53)
1998	3115	2421 (78)	553	184 (33)	2152	1089 (51)
1999	3988	3143 (79)	652	192 (29)	3587	1781 (50)
2000	5421	4263 (79)	795	237 (30)	5736	2950 (51)
2001	6797	5523 (81)	830	256 (31)	8294	4607 (56)
2002	8163	6873 (84)	898	310 (35)	11797	7296 (62)
2003	9232	8094 (88)	954	356 (37)	15878	10925 (69)
2004	10612	9801 (92)	1007	450 (45)	21415	16052 (75)
2005	11883	11307 (95)	1064	555 (52)	26398	20590 (78)
2006	12831	12358 (96)	1068	637 (60)	30977	24772 (80)
2007	13660	13186 (97)	1085	742 (68)	34751	28604 (82)
2008	14367	13906 (97)	1124	813 (72)	36781	30522 (83)
2009	15109	14648 (97)	1133	854 (75)	38942	32144 (83)
2010	15508	15019 (97)	1160	872 (75)	39499	32264 (82)
2011	15732	15224 (97)	1156	898 (78)	38959	31861 (82)

Appendix Table A3: Completeness of BMI data in UK databases.

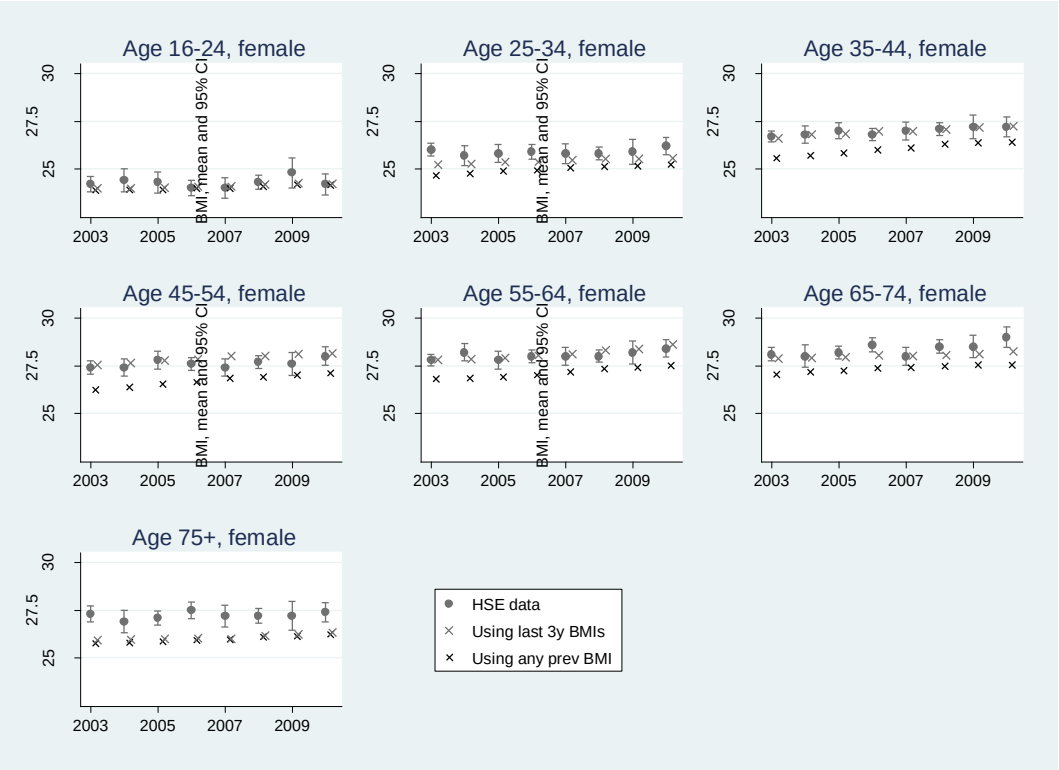
Database	**Calendar time-span**	**Completeness Information**
CPRD (current study)	2005-2011	51% of patients had a BMI recorded in the last 3 years. 62% of newly registered patients had a BMI recorded by 12 months after registration in our study.
QResearch [21]	2001-2007	By March 2007, 58% of registered patients had their BMI recorded in the past 5 years.
THIN [22]	2004-2006	62% of patients had a height recording and 66% had a weight recording within 12 months of registration.

Appendix Figure A1: Mean BMI over calendar time comparing those with BMI recorded in CPRD (English practices) with the Health Survey for England 2010 data, stratified by gender and age group

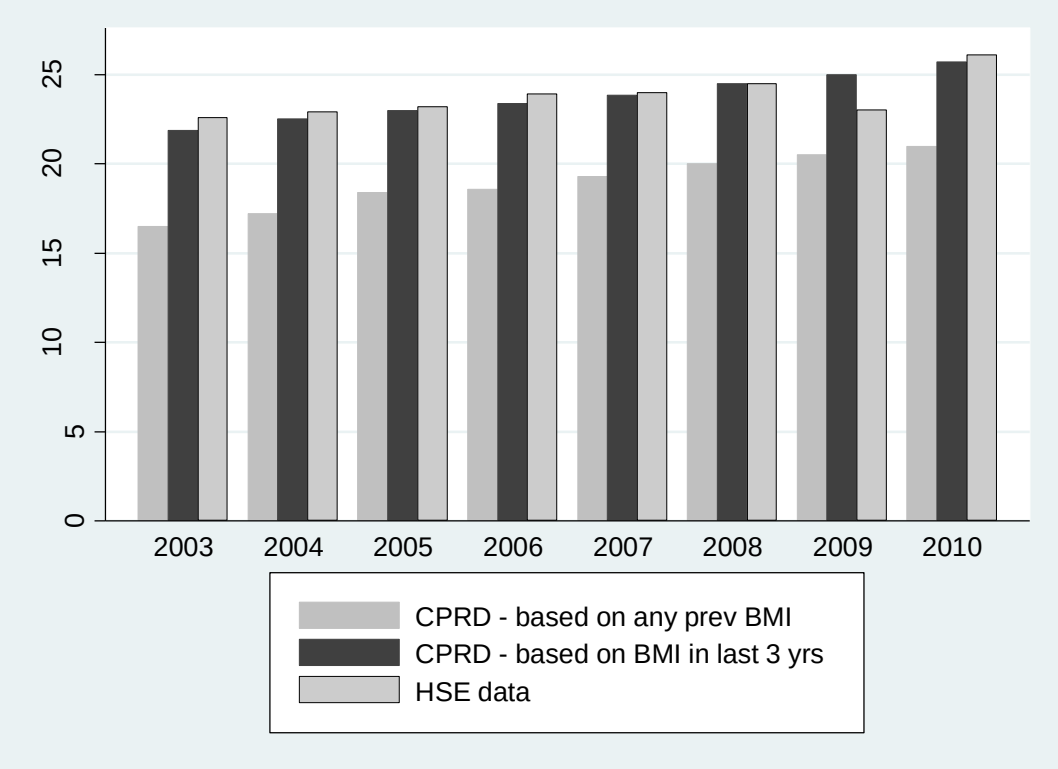
(a) Men



(b) Women

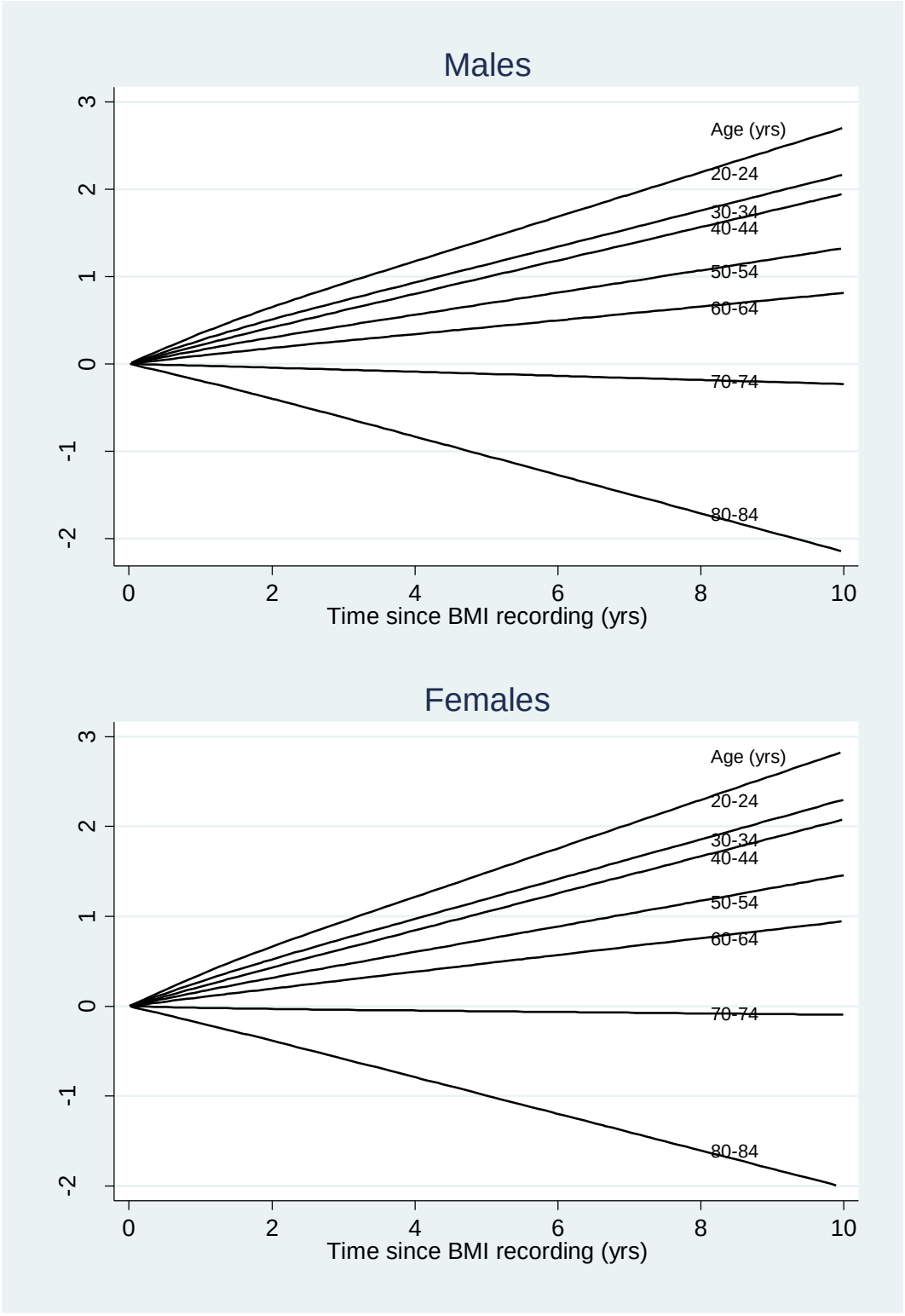


Appendix Figure A2: Percentage of individuals in CPRD (English practices) classified as obese (BMI≥30kg/m^2) compared with Health Survey for England data



Notes:
Proportions are age- and sex-standardised to the HSE population structure

Appendix Figure A3: Modelled change in BMI over time since BMI recording, by age and sex



Notes:
- Based on a weighted least squares regression model for change in BMI with a 3 knot spline for time since BMI and interaction terms between spline terms and age/sex indicator variables (generating a separate modelled curve for each age/sex stratum). No constant term was included, to ensure that predicted change in BMI at time zero would be zero.
- For clarity, only selected age groups are shown in the Figure; the model also included the age groups 16-19, 25-29, 35-39, 45-49, 55-59, 65-69, 75-79, 85-89, 90+ years.