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

Journal:	*BMJ Open*
Manuscript ID:	bmjopen-2013-003389
Article Type:	Research
Date Submitted by the Author:	11-Jun-2013
Complete List of Authors:	Bhaskaran, Krishnan; LSHTM, NCDE Forbes, Harriet; LSHTM, NCDE Douglas, Ian; LSHTM, NCDE Leon, David; LSHTM, NCDE Smeeth, Liam; LSHTM, NCDE
Primary Subject Heading:	Epidemiology
Secondary Subject Heading:	Epidemiology, Research methods
Keywords:	EPIDEMIOLOGY, PRIMARY CARE, STATISTICS & RESEARCH METHODS

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

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

Faculty of Epidemiology & Population Health, London School of Hygiene and Tropical Medicine, Keppel Street, London, WC1E 7HT, UK

Corresponding Author: Krishnan Bhaskaran, London School of Hygiene and Tropical Medicine, Keppel Street, London, WC1E 7HT. Email: krishnan.bhaskaran@lshtm.ac.uk. Tel: + 44 (0) 20 7927 2268

Key words: CPRD, BMI, missing data, primary care databases, obesity.

Word Count: 3383

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 Clinical Practice Research Datalink (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 (2003-10) were age- 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 1990-94 to 77% in 2005-11, 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 0.75-1.1kg/m^2. Restricting to those with a recent (≤3 years) BMI resulted in mean BMI estimates closer to HSE (≤0.28kg/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 (≤0.37kg/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

Article focus:

- Body mass index (BMI) data are frequently used in epidemiological analyses of primary care databases such as the UK Clinical Practice Research Datalink (CPRD), however their completeness and representativeness have not previously been assessed in detail.
- The aim of this article is to provide information on the completeness of BMI in CPRD primary care data, on their representativeness, and on the implications for their practical use in research.

Key messages:

- We found that completeness of BMI recordings in the Clinical Practice Research Datalink increased from 37% in 1990-4 to 77% in 2005-11 and differed by age and sex.
- At specific calendar time points, the most recent BMI recorded for a large proportion of patients was over 3 years old and was unlikely to reflect current BMI.
- The optimal strategy for assigning BMI status is likely to depend on the specific study population and research context. We suggest one possible approach that uses a model-based imputation of current BMI to reduce BMI misclassification.

Strengths and limitations of this study:

- Results presented here are based on a large random sample from the CPRD, therefore we can confidently generalise the findings to the whole CPRD database, and to similar databases based on UK primary care records.
- To assess the representativeness of CPRD 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 both developed and developing countries[3, 4] and is a growing concern to 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]. Because of 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 medical records for approximately 8% of the UK population.[6] However, a shortcoming of these databases is that lifestyle data, such as BMI, tend to be opportunistically recorded 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.

Methods

Data sources

Clinical Practice Research Datalink (CPRD)

The Clinical Practice Research Datalink (CPRD) is a clinical database comprising anonymised computerised medical records from general practitioners (GPs) in the United Kingdom. Approximately 8% of the UK population are currently included and the database is broadly representative of the UK population.[15] CPRD contains 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.

Body mass 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). Patient 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.[16] The HSE is an annual survey designed to produce a representative sample of the adult population aged ≥16 years and living in private households. The methods are described in detail elsewhere.[17] Surveys were interviewer

administered with interviewers measuring the weight and height of all participants. Data from 2003-10 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 Socio-economic Classification system).

Statistical methods

Completeness of BMI data in CPRD

In the main analyses BMI completeness data in CPRD were estimated by calendar period (1990-4, 1995-9, 2000-4, 2005-11). 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 mid-point 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 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 BMI category. We have not presented confidence intervals for these descriptive statistics because the sample size made sampling error negligible (for example, the standard errors for the proportions with complete BMI data in age and calendar year subgroups were all <0.5%).

Comparison of CPRD BMI data with Health Survey for England 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, HSE data were adjusted for non-response). CPRD mean BMI was based on patients registered and under up-to-standard follow-up at the mid-point of the calendar year. We produced two sets of CPRD mean BMI statistics: firstly we used last BMI observation carried forward (regardless of how long ago recorded); secondly we restricted to patients with a recent BMI available (up to 3 years before the mid-point of the calendar year). As above, confidence intervals are not presented because there was negligible sampling error (maximum standard error=0.02kg/m²). To make like-with-like comparisons with HSE, CPRD data were restricted to English practices, and mean BMI was age- and sex-standardised to the HSE population structure Proportions classified as obese (BMI≥30kg/m^2) over time 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 BMI recording (restricting to individuals with BMI records ≤ 10 years apart). We fitted a linear regression model with change in BMI as the outcome, and elapsed time included as a 3 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 heteroskedasticity.[18]

Having specified a model for change in BMI over time, we first explored its performance among individuals with at least 2 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 from 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 included all individuals with a BMI record in the previous 10 years and used the model described above to impute current BMI at the mid-point 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 5 imputations) to account for uncertainty in the modelled changes over time.[19]

The study was approved by the London School of Hygiene and Tropical Medicine Ethics Committee.

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 2005-11(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 2005-11). 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 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 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 (Appendix Table A2). This is in line with Quality and Outcomes Framework (QOF) which has encouraged BMI monitoring in these

conditions since 2004.[20] 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, out of 75518 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 (7675/250430 = 3.0% in the most recent period).

Summary statistics using complete CPRD BMI data and comparison with Health Survey for England

We found that age- and sex-standardised mean BMI based on CPRD data was consistently and substantially lower (by up to 1.1kg/m^2) than in the HSE data (mean BMI in CPRD = 25.7kg/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 to 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 CPRD data underestimating the HSE statistics by only between 0.04 to 0.28kg/m^2 in individual calendar years, and the CPRD estimate falling within the HSE confidence interval for 2 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- and sex-stratified data demonstrated similar patterns, except that in the eldest age group (75+ years), restricting to those with recent BMI measures did not bring the estimated BMI substantially closer to HSE figures (Appendix Figure A1).

We also compared the proportions classified as obese between CPRD and HSE (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 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 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 (Appendix Figure A3). We tested the predictive performance of our model by predicting the most recent BMI based on the previous one, among CPRD patients 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 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 = -0.70kg/m^2 [IQR -2.18 to +0.56] using last observation carried forward, compared with +0.11kg/m^2 [-1.29 to +1.40] using 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. Estimated mean BMIs were now in line with those based on only recent data in the earlier analysis, and were only between 0.04 and 0.37kg/m^2 lower than HSE statistics in individual calendar years (Figure 3, circles). Even with multiple imputation, confidence intervals remained extremely narrow (<0.07kg/m^2) due to the large sample size, so 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 to 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 1990-94 to 77% in 2005-11). 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 equivalent HSE estimates (by up to 1.1kg/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, mean BMI statistics were closer to 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 (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);[21] 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 2005-11). As in our study, the QRESEARCH report showed an increase in completeness over time, rising from 42% in 2000/01 to 58% in 2007. In a third UK primary care database, THIN (The Health Improvement Network), 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.[22]

Explanation of findings

Completeness

Increasing 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 [23, 24] and different prevalence's of diseases in which it is important to monitor BMI.

Comparison of CPRD BMI data with Health Survey for England data

Mean BMI based on the CPRD population was lower in each calendar year than 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 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 [25-27]. A simple adjustment of outdated BMIs based on these modelled changes over time brought CPRD mean BMI statistics in line with 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 (for example, 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.[22]

Second, our analyses suggest that the common practice of assigning BMI status based on the nearest/most recent 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 numbers 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 done 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. Ultimately, the importance of these issues and the optimal strategy to use is likely to depend on the particular study and the characteristics of the study population.

Strengths and Limitations

Results presented here are based on a large random sample from the CPRD, therefore we can confidently generalise the findings to the whole CPRD database. Although we cannot assume these findings will relate to other routinely collected primary care databases in UK based on other IT systems (CPRD is based on practices using VISION), they are likely to be similar. This study did not look at BMI recordings among children as this would require a different strategy. Completeness among 16-24 year 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 (sample size 14,836 in 2003 and 8,420 in 2010), utilising height and weight recordings measured by a trained interviewer, and is weighted for non-response.[17, 28] 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),[17] which if related to BMI status, may bias the estimates of mean BMI in HSE.

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. 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.

Conflicts of interest

The authors declare no conflicts of interest.

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.

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: 2009.
2. Flegal KM, Graubard BI, Williamson DF, Gail MH. Excess deaths associated with underweight, overweight, and obesity. Jama-J Am Med Assoc 2005;293(15):1861-7.
3. Swinburn BA, Sacks G, Hall KD, McPherson K, Finegood DT, Moodie ML, et al. Obesity 1 The global obesity pandemic: shaped by global drivers and local environments. Lancet 2011;378(9793):804-14.
4. Kelly T, Yang W, Chen CS, Reynolds K, He J. Global burden of obesity in 2005 and projections to 2030. Int J Obesity 2008;32(9):1431-7.
5. NHS Information Centre. Health survey for England - 2010: health and lifestyles 2011. Available from: http://www.ic.nhs.uk/pubs/hse10report.
6. CPRD. Clinical Practice Research Database (CPRD) website. Available from: http://www.cprd.com/intro.asp.
7. Delaney JA, Daskalopoulou SS, Brophy JM, Steele RJ, Opatrny L, Suissa S. 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, Watson J, Wise L, Beral V. Oral bisphosphonates and risk of cancer of oesophagus, stomach, and colorectum: case-control analysis within a UK primary care cohort. Brit Med J 2010;341.
9. Tzoulaki I, Molokhia M, Curcin V, Little MP, Millett CJ, Ng A, 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. Brit Med J 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(1):116-20.
11. Andersohn F, Schade R, Suissa S, Garbe E. Long-Term Use of Antidepressants for Depressive Disorders and the Risk of Diabetes Mellitus. Am J Psychiat 2009;166(5):591-8.
12. Lawrenson R, Todd JC, Leydon GM, Williams TJ, Farmer RDT. Validation of the diagnosis of venous thromboembolism in general practice database studies. Br J Clin Pharmacol 2000;49(6):591-6.
13. Jick H, Zornberg GL, Jick SS, Seshadri S, Drachman DA. Statins and the risk of dementia. Lancet 2000;356(9242):1627-31.
14. van Staa TP, Wegman S, de Vries F, Leufkens B, Cooper C. Use of statins and risk of fractures. Jama-J Am Med Assoc 2001;285(14):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. Series MB6 No. 2, editor. London: Office for National Statistics; 2000.
16. NHS Information Centre. Health Survey for England - 2010: Trend tables: National Health Service; [cited on 19th June 2012]. Available from: 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.
17. Aresu M, Boodhna G, Bryson A, Bridges S, Chaudhury M, Craig R, et al. Volume 2: Methods and documentation. In: Craig R, Mindell J, editors. Health Survey for England 2010. Leeds: NHS Information Centre for health and social care; 2011.
18. Greene WH. Econometric Analysis. Upper Saddle River, New Jersey: Prentice Hall.1997.

19. Rubin DB. Multiple Imputation for Nonresponse in Surveys: J. Wiley & Sons, New York. ; 1987.
20. NHS. Quality and Outcomes Framework guidance for GMS contract 2011/12. Employers and British Medical Association; 2011.
21. NHS Information Centre. A summary of public health indicators using electronic data from primary care 2008 [cited on 17th December 2012]. Available from: 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.
22. Marston L, Carpenter JR, Walters KR, Morris RW, Nazareth I, Petersen I. Issues in multiple imputation of missing data for large general practice clinical databases. Pharmacoepidem Dr S 2010;19(6):618-26.
23. Rowlands S, Moser K. Consultation rates from the General Practice Research Database. Br J Gen Pract 2002;52(481):658-60.
24. The Health and Social Care Information Centre. Trends in Consultation Rates in General Practice 1995 to 2009 2009 [cited on 17th December 2012]. Available from: http://www.ic.nhs.uk/article/2021/Website-Search?productid=729&q=qresearch&sort=Relevance&size=10&page=1&area=both#top.
25. 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(6):1215-23.
26. Silverwood RJ, Pierce M, Thomas C, Hardy R, Ferro C, Sattar N, et al. Overweight across adult life and kidney function at age 60-4 years: the 1946 British birth cohort study. (awaiting publication) 2012.
27. Tirosh A, Shai I, Afek A, Dubnov-Raz G, Ayalon N, Gordon B, et al. Adolescent BMI Trajectory and Risk of Diabetes versus Coronary Disease. N Engl J Med 2011;364(14):1315-25.
28. Mindell J, Biddulph JP, Hirani V, Stamatakis E, Craig R, Nunn S, et al. Cohort Profile: The Health Survey for England. Int J Epidemiol 2012:1-9.

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

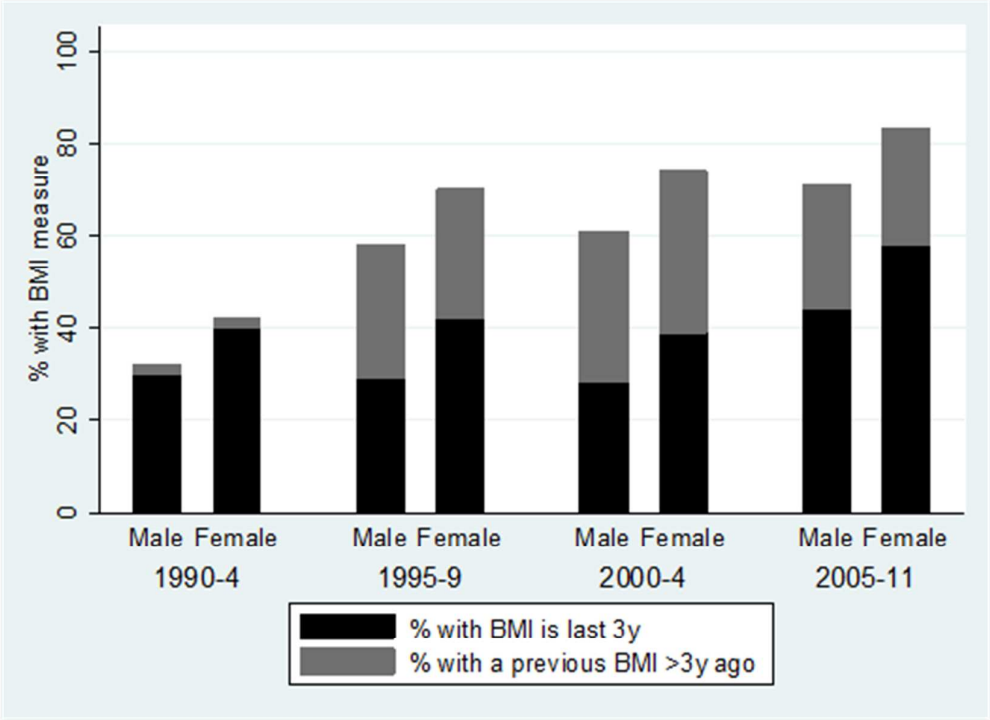
Age group (yrs)	1990-4	1995-9	2000-4	2005-2011
16-24[a]				
N registered	11423	17501	34452	42546
BMI in previous 3y (%)	26	28	25	32
BMI any previous (%)	26	37	30	37
25-34				
N registered	17477	29923	48659	50413
BMI in previous 3y (%)	37	39	36	49
BMI any previous (%)	38	66	67	72
35-44				
N registered	15953	28838	55991	61014
BMI in previous 3y (%)	36	36	31	46
BMI any previous (%)	39	67	71	80
45-54				
N registered	14507	27765	48093	55564
BMI in previous 3y (%)	39	37	32	50
BMI any previous (%)	42	70	73	84
55-64				
N registered	11680	20843	42258	49380
BMI in previous 3y (%)	42	40	37	57
BMI any previous (%)	44	74	77	87
65-74				
N registered	10678	17605	30997	34508
BMI in previous 3y (%)	36	37	40	67
BMI any previous (%)	38	71	79	91
75+				
N registered	8637	16005	29384	32523
BMI in previous 3y (%)	28	32	37	64
BMI any previous (%)	28	56	69	87
Total				
N registered	90355	158480	289834	325948
BMI in previous 3y (%)	35	36	34	51
BMI any previous (%)	37	64	67	77

N registered is all those under follow-up at mid-point of the period

[a]Note, BMI measurements from age <16 years were not counted in this analysis, hence completeness in the 16-24 age group may be artificially low

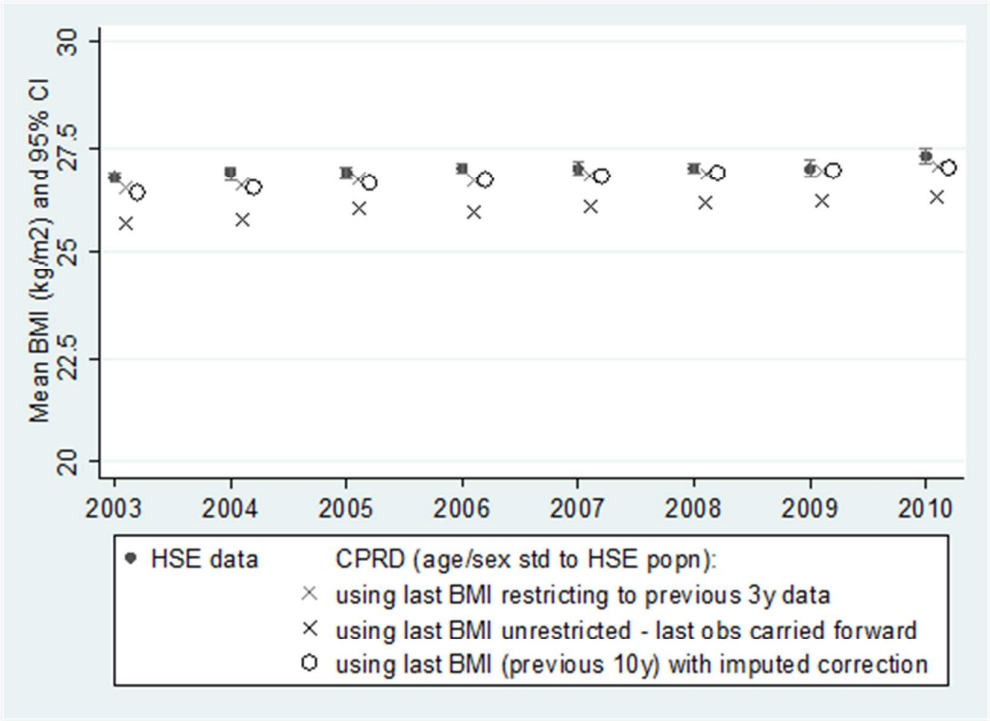
Figure 1: Initial data processing to generate BMI for analysis

Figure 2: Completeness of BMI data in CPRD, by gender and calendar period



Note: Completeness data for each calendar period are based on all those under follow-up at mid-point of the period

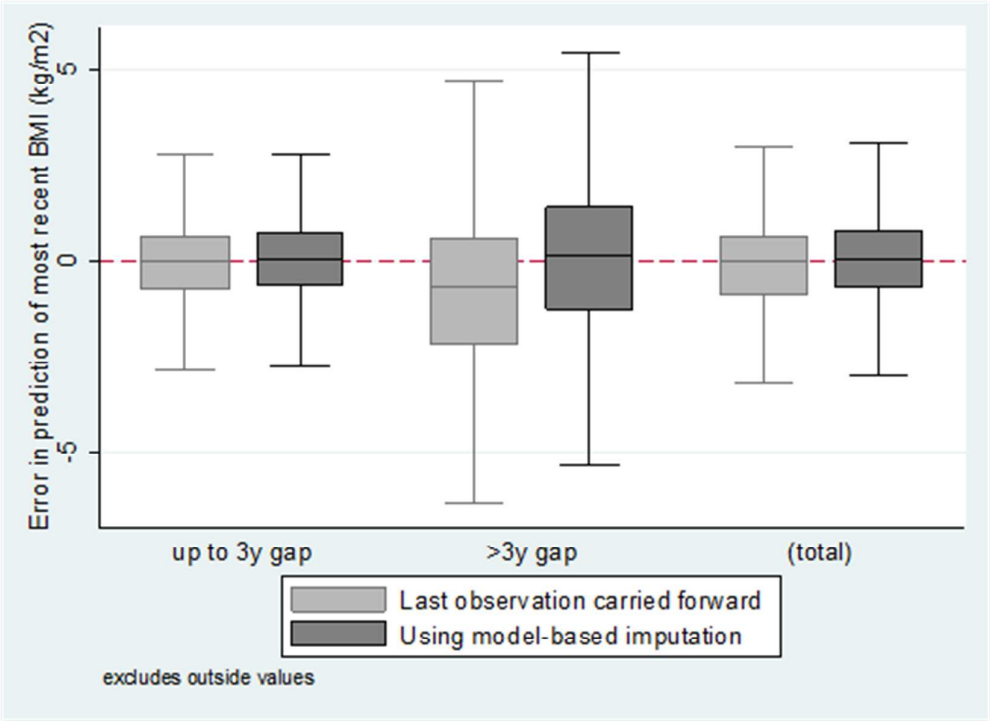
Figure 3: Mean BMI over calendar time comparing those with BMI recorded in CPRD (English practices) with the Health Survey for England 2010 data



Note: CPRD figures are age- and sex- standardised to the Health Survey for England study population

CPRD statistics are based on all patients registered at the mid-point of the calendar period and with a suitable previous BMI measure available (i.e. either any previous, or within the last 3 years)

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



Author contributions

I, Krishnan Bhaskaran, developed the analytical strategy for this paper, processed and analysed the data and wrote the paper.

I, Harriet Forbes, was involved in discussing the data processing and analysis of the data, as well as the writing of the paper.

I, Liam Smeeth, was involved in discussions of the analytical approach to this study and made comments on the analysis and the writing of the paper.

I, Ian Douglas, was involved in discussions of the analytical approach to this study and made comments on the analysis and the writing of the paper.

I, David Leon, was involved in discussions of the analytical approach to this study and made comments on the analysis and the writing of the paper.

STROBE Statement—Checklist of items that should be included in reports of ***cohort studies***

	Item No	Recommendation
Title and abstract	1	(*a*) Indicate the study's design with a commonly used term in the title or the abstract n/a (we did not think there was an appropriate design keyword/term to describe this study as it is not a standard "exposure/outcome" study but is rather providing data quality information on a common exposure/covariate)
		(*b*) Provide in the abstract an informative and balanced summary of what was done and what was found P2
Introduction		
Background/rationale	2	Explain the scientific background and rationale for the investigation being reported P4
Objectives	3	State specific objectives, including any prespecified hypotheses P4
Methods		
Study design	4	Present key elements of study design early in the paper P6-7
Setting	5	Describe the setting, locations, and relevant dates, including periods of recruitment, exposure, follow-up, and data collection P5-6
Participants	6	(*a*) Give the eligibility criteria, and the sources and methods of selection of participants. Describe methods of follow-up P5-6
		(*b*) For matched studies, give matching criteria and number of exposed and unexposed
Variables	7	Clearly define all outcomes, exposures, predictors, potential confounders, and effect modifiers. Give diagnostic criteria, if applicable P5-7
Data sources/ measurement	8*	For each variable of interest, give sources of data and details of methods of assessment (measurement). Describe comparability of assessment methods if there is more than one group P5-6
Bias	9	Describe any efforts to address potential sources of bias P6-7
Study size	10	Explain how the study size was arrived at P5
Quantitative variables	11	Explain how quantitative variables were handled in the analyses. If applicable, describe which groupings were chosen and why P6-7
Statistical methods	12	(*a*) Describe all statistical methods, including those used to control for confounding P6-7
		(*b*) Describe any methods used to examine subgroups and interactions P6-7
		(*c*) Explain how missing data were addressed P7
		(*d*) If applicable, explain how loss to follow-up was addressed n/a
		(*e*) Describe any sensitivity analyses

		n/a
Results		
Participants	13*	(a) Report numbers of individuals at each stage of study—eg numbers potentially eligible, examined for eligibility, confirmed eligible, included in the study, completing follow-up, and analysed
		(b) Give reasons for non-participation at each stage
		(c) Consider use of a flow diagram FIG 1
Descriptive data	14*	(a) Give characteristics of study participants (eg demographic, clinical, social) and information on exposures and potential confounders
		(b) Indicate number of participants with missing data for each variable of interest
		(c) Summarise follow-up time (eg, average and total amount) P8-9 and FIG 2
Outcome data	15*	Report numbers of outcome events or summary measures over time n/a (no specific outcome)
Main results	16	(*a*) Give unadjusted estimates and, if applicable, confounder-adjusted estimates and their precision (eg, 95% confidence interval). Make clear which confounders were adjusted for and why they were included
		(*b*) Report category boundaries when continuous variables were categorized
		(*c*) If relevant, consider translating estimates of relative risk into absolute risk for a meaningful time period n/a (not an "exposure/outcome" study)
Other analyses	17	Report other analyses done—eg analyses of subgroups and interactions, and sensitivity analyses P9-11
Discussion		
Key results	18	Summarise key results with reference to study objectives P11-12
Limitations	19	Discuss limitations of the study, taking into account sources of potential bias or imprecision. Discuss both direction and magnitude of any potential bias P14-15
Interpretation	20	Give a cautious overall interpretation of results considering objectives, limitations, multiplicity of analyses, results from similar studies, and other relevant evidence P15
Generalisability	21	Discuss the generalisability (external validity) of the study results P14
Other information		
Funding	22	Give the source of funding and the role of the funders for the present study and, if applicable, for the original study on which the present article is based P16

*Give information separately for exposed and unexposed groups.

Note: An Explanation and Elaboration article discusses each checklist item and gives methodological background and published examples of transparent reporting. The STROBE checklist is best used in conjunction with this article (freely available on the Web sites of PLoS Medicine at http://www.plosmedicine.org/, Annals of Internal Medicine at http://www.annals.org/, and Epidemiology at http://www.epidem.com/). Information on the STROBE Initiative is available at http://www.strobe-statement.org.

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

Journal:	*BMJ Open*
Manuscript ID:	bmjopen-2013-003389.R1
Article Type:	Research
Date Submitted by the Author:	06-Aug-2013
Complete List of Authors:	Bhaskaran, Krishnan; LSHTM, NCDE Forbes, Harriet; LSHTM, NCDE Douglas, Ian; LSHTM, NCDE Leon, David; LSHTM, NCDE Smeeth, Liam; LSHTM, NCDE
Primary Subject Heading:	Epidemiology
Secondary Subject Heading:	Epidemiology, Research methods
Keywords:	EPIDEMIOLOGY, PRIMARY CARE, STATISTICS & RESEARCH METHODS

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

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

Faculty of Epidemiology & Population Health, London School of Hygiene and Tropical Medicine, Keppel Street, London, WC1E 7HT, UK

Corresponding Author: Krishnan Bhaskaran, London School of Hygiene and Tropical Medicine, Keppel Street, London, WC1E 7HT. Email: krishnan.bhaskaran@lshtm.ac.uk. Tel: + 44 (0) 20 7927 2268

Key words: CPRD, BMI, missing data, primary care databases, obesity.

Word Count: 3680

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 Clinical Practice Research Datalink (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 (2003-10) were age- 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 1990-94 to 77% in 2005-11, 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 0.75-1.1kg/m^2. Restricting to those with a recent (≤3 years) BMI resulted in mean BMI estimates closer to HSE (≤0.28kg/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 (≤0.37kg/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

Article focus:

- Body mass index (BMI) data are frequently used in epidemiological analyses of primary care databases such as the UK Clinical Practice Research Datalink (CPRD), however their completeness and representativeness have not previously been assessed in detail.
- The aim of this article is to provide information on the completeness of BMI in CPRD primary care data, on their representativeness, and on the implications for their practical use in research.

Key messages:

- We found that completeness of BMI recordings in the Clinical Practice Research Datalink increased from 37% in 1990-4 to 77% in 2005-11 and differed by age and sex.
- At specific calendar time points, the most recent BMI recorded for a large proportion of patients was over 3 years old and was unlikely to reflect current BMI.
- The optimal strategy for assigning BMI status is likely to depend on the specific study population and research context. We suggest one possible approach that uses a model-based imputation of current BMI to reduce BMI misclassification.

Strengths and limitations of this study:

- Results presented here are based on a large random sample from the CPRD, therefore we can confidently generalise the findings to the whole CPRD database, and to similar databases based on UK primary care records.
- To assess the representativeness of CPRD 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 both developed and developing countries[3 4] and is a growing concern to 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]. Because of 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 (i.e. recorded when the patient is attending for other reasons, or when 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)

The Clinical Practice Research Datalink (CPRD) is a clinical database comprising anonymised computerised medical records from general practitioners (GPs) in the United Kingdom. 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 a 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.

Body mass 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] The 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 8,420 in 2010),. Surveys were interviewer administered with interviewers measuring the weight and height of all participants. Data from 2003-10 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 Socio-economic 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 (1990-4, 1995-9, 2000-4, 2005-11). 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 mid-point 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 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

BMI category. We have not presented confidence intervals for these descriptive statistics because the sample size made sampling error negligible (for example, the standard errors for the proportions with complete BMI data in age and calendar year subgroups were all <0.5%).

Comparison of CPRD BMI data with Health Survey for England 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, HSE data were adjusted for non-response). CPRD mean BMI was based on patients registered and under up-to-standard follow-up at the mid-point of the calendar year. We produced two sets of CPRD mean BMI statistics: firstly we used last BMI observation carried forward (regardless of how long ago recorded); secondly we restricted to patients with a recent BMI available (up to 3 years before the mid-point of the calendar year). As above, confidence intervals are not presented because there was negligible sampling error (maximum standard error=0.02kg/m²). 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- and sex-standardised to the HSE population structure Proportions classified as obese (BMI≥30kg/m^2) over time 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 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 wasincluded as a 3 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 heteroskedasticity.[20]

Having specified a model for change in BMI over time, we first explored its performance among individuals with at least 2 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 from 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 included all individuals with a BMI record in the previous 10 years and used the model described above to impute current BMI at the mid-point 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 5 imputations) to account for uncertainty in the modelled changes over time.[21]

The study was approved by the London School of Hygiene and Tropical Medicine Ethics Committee.

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 2005-11(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 2005-11). 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 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 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 (Appendix Table A2). This is in line with 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, out of 75518 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 (7675/250430 = 3.0% in the most recent period).

Summary statistics using complete CPRD BMI data and comparison with Health Survey for England

We found that age- and sex-standardised mean BMI based on CPRD data was consistently and substantially lower (by up to 1.1kg/m^2) than in the HSE data (mean BMI in CPRD = 25.7kg/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 to 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 CPRD data underestimating the HSE statistics by only between 0.04 to 0.28kg/m^2 in individual calendar years, and the CPRD estimate falling within the HSE confidence interval for 2 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- and sex-stratified data demonstrated similar patterns, except that in the eldest age group (75+ years), restricting to those with recent BMI measures did not bring the estimated BMI substantially closer to HSE figures (Appendix Figure A1).

We also compared the proportions classified as obese between CPRD and HSE (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 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 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 (Appendix Figure A3). We tested the predictive performance of our model by predicting the most recent BMI based on the previous one, among CPRD patients 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 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 = -0.70kg/m^2 [IQR -2.18 to +0.56] using last observation carried forward, compared with +0.11kg/m^2 [-1.29 to +1.40] using 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. Estimated mean BMIs were now in line with those based on only recent data in the earlier analysis, and were only between 0.04 and 0.37kg/m^2 lower than HSE statistics in individual calendar years (Figure 3, circles). Even with multiple imputation, confidence intervals remained extremely narrow (<0.07kg/m^2) due to the large sample size, so 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 to 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 1990-94 to 77% in 2005-11). 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 equivalent HSE estimates (by up to 1.1kg/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, mean BMI statistics were closer to 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 (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 2005-11). As in our study, the QRESEARCH report showed an increase in completeness over time, rising from 42% in 2000/01 to 58% in 2007. In a third UK primary care database, THIN (The Health Improvement Network), 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 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 Health Survey for England data

Mean BMI based on the CPRD population was lower in each calendar year than 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 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 CPRD mean BMI statistics in line with 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 (for example, 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 recent 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 numbers 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 done 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 two-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, and the optimal strategy to use is likely to depend on the particular study and the characteristics of the study population.

Strengths and Limitations

Results presented here are based on a large random sample from the CPRD, therefore we can confidently generalise the findings to the whole CPRD database. Although we cannot assume these findings will relate to other routinely collected primary care databases in UK based on other IT systems (CPRD is based on practices using VISION), they are likely to be similar. This study did not look at BMI recordings among children as this would require a different strategy. Completeness among 16-24 year 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 focussed on the mean (and in the appendix, on the proportion classed as obese); these are the principal statistics published in 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.

Conflicts of interest

The authors declare no conflicts of interest.

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.

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. In: Organisation WH, ed. Geneva, Switzerland, 2009.
2. Flegal KM, Graubard BI, Williamson DF, et al. Excess deaths associated with underweight, overweight, and obesity. Jama-J Am Med Assoc 2005;**293**(15):1861-67
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**(9793):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**(9):1431-37 doi: Doi 10.1038/Ijo.2008.102[published Online First: Epub Date]|.
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 doi: 1471-2261-7-38 [pii]

10.1186/1471-2261-7-38[published Online First: Epub Date]|.
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 doi: 10.1136/bmj.c4444[published Online First: Epub Date]|.
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. British Medical Journal 2009;**339**:b4731 doi: Artn B4731

Doi 10.1136/Bmj.B4731[published Online First: Epub Date]|.
10. Douglas I, Smeeth L, Irvine D. The use of antidepressants and the risk of haemorrhagic stroke: a nested case control study. Brit J Clin Pharmaco 2011;**71**(1):116-20 doi: DOI 10.1111/j.1365-2125.2010.03797.x[published Online First: Epub Date]|.
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 Psychiat 2009;**166**(5):591-98 doi: DOI 10.1176/appi.ajp.2008.08071065[published Online First: Epub Date]|.
12. Lawrenson R, Todd JC, Leydon GM, et al. Validation of the diagnosis of venous thromboembolism in general practice database studies. Brit J Clin Pharmaco 2000;**49**(6):591-96
13. Jick H, Zornberg GL, Jick SS, et al. Statins and the risk of dementia. Lancet 2000;**356**(9242):1627-31
14. van Staa TP, Wegman S, de Vries F, et al. Use of statins and risk of fractures. Jama-J Am Med Assoc 2001;**285**(14):1850-55
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, 2007:341-48.
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**(4):588-96 doi: 10.1111/j.1365-2125.2010.03733.x[published Online First: Epub Date]|.
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, New Jersey: Prentice Hall., 1997.
21. Rubin DB. *Multiple Imputation for Nonresponse in Surveys*: J. Wiley & Sons, New York. , 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**(6):618-26 doi: 10.1002/pds.1934[published Online First: Epub Date]|.
25. Rowlands S, Moser K. Consultation rates from the General Practice Research Database. Brit J Gen Pract 2002;**52**(481):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. Journal of Hypertension 2007;**25**(6):1215-23
28. Silverwood RJ, Pierce M, Thomas C, et al. Overweight across adult life and kidney function at age 60-4 years: the 1946 British birth cohort study. (awaiting publication) 2012
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**(14):1315-25 doi: 10.1056/NEJMoa1006992[published Online First: Epub Date]|.
30. Welch C, Bartlett J, Petersen I. Application of multiple imputation using the two-fold fully conditional specification algorithm in longitudinal clinical data. (In Press). Stata Journal 2013
31. Mindell J, Biddulph JP, Hirani V, et al. Cohort Profile: The Health Survey for England. International Journal of Epidemiology 2012:1-9 doi: 10.1093/ije/dyr199[published Online First: Epub Date]|.

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

Age group (yrs)	1990-4	1995-9	2000-4	2005-2011
16-24[a]				
N registered	11423	17501	34452	42546
BMI in previous 3y (%)	26	28	25	32
BMI any previous (%)	26	37	30	37
25-34				
N registered	17477	29923	48659	50413
BMI in previous 3y (%)	37	39	36	49
BMI any previous (%)	38	66	67	72
35-44				
N registered	15953	28838	55991	61014
BMI in previous 3y (%)	36	36	31	46
BMI any previous (%)	39	67	71	80
45-54				
N registered	14507	27765	48093	55564
BMI in previous 3y (%)	39	37	32	50
BMI any previous (%)	42	70	73	84
55-64				
N registered	11680	20843	42258	49380
BMI in previous 3y (%)	42	40	37	57
BMI any previous (%)	44	74	77	87
65-74				
N registered	10678	17605	30997	34508
BMI in previous 3y (%)	36	37	40	67
BMI any previous (%)	38	71	79	91
75+				
N registered	8637	16005	29384	32523
BMI in previous 3y (%)	28	32	37	64
BMI any previous (%)	28	56	69	87
Total				
N registered	90355	158480	289834	325948
BMI in previous 3y (%)	35	36	34	51
BMI any previous (%)	37	64	67	77

N registered is all those under follow-up at mid-point of the period

[a]Note, BMI measurements from age <16 years were not counted in this analysis, hence completeness in the 16-24 age group may be artificially low

Figure 1: Initial data processing to generate BMI for analysis

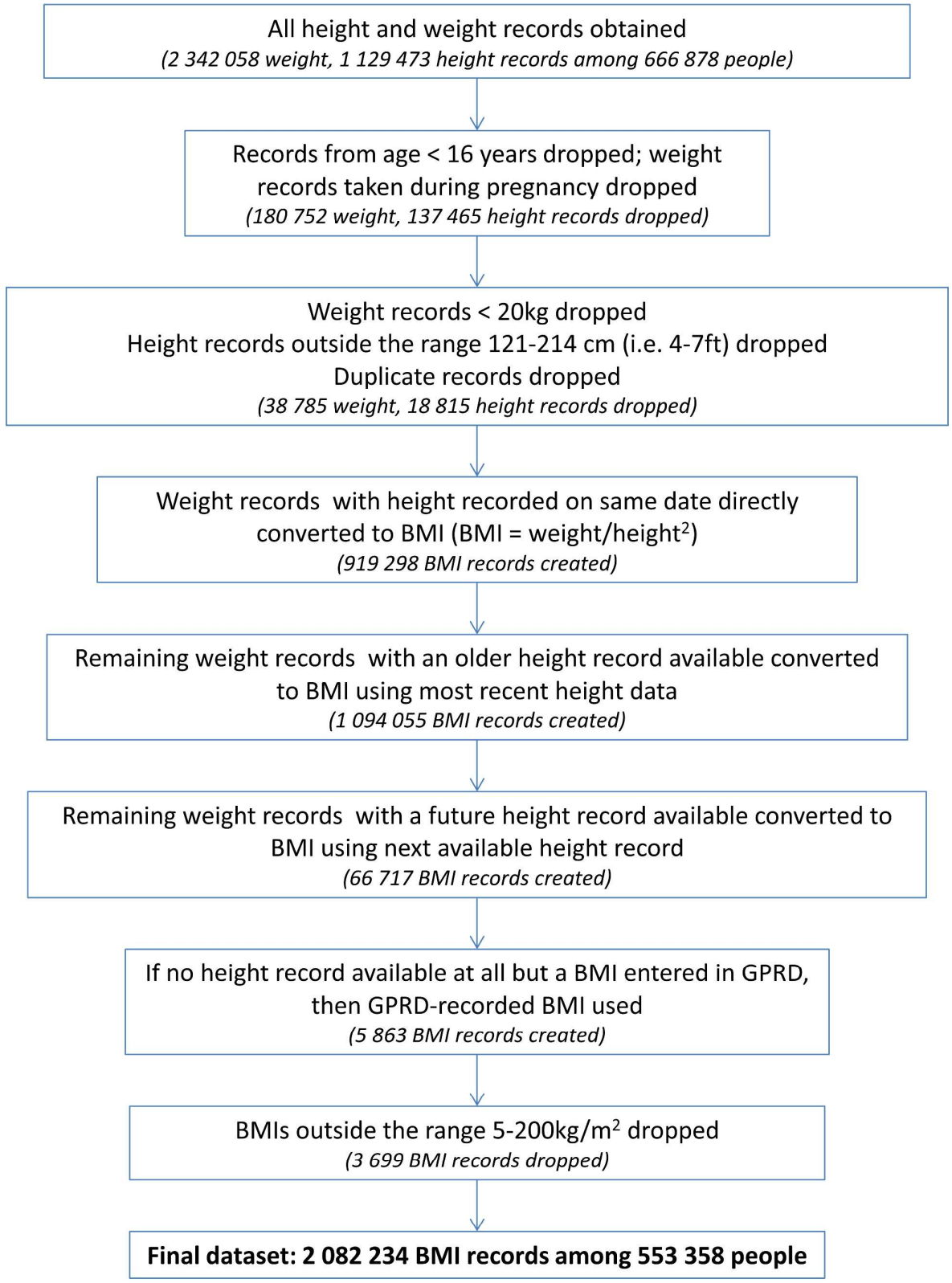
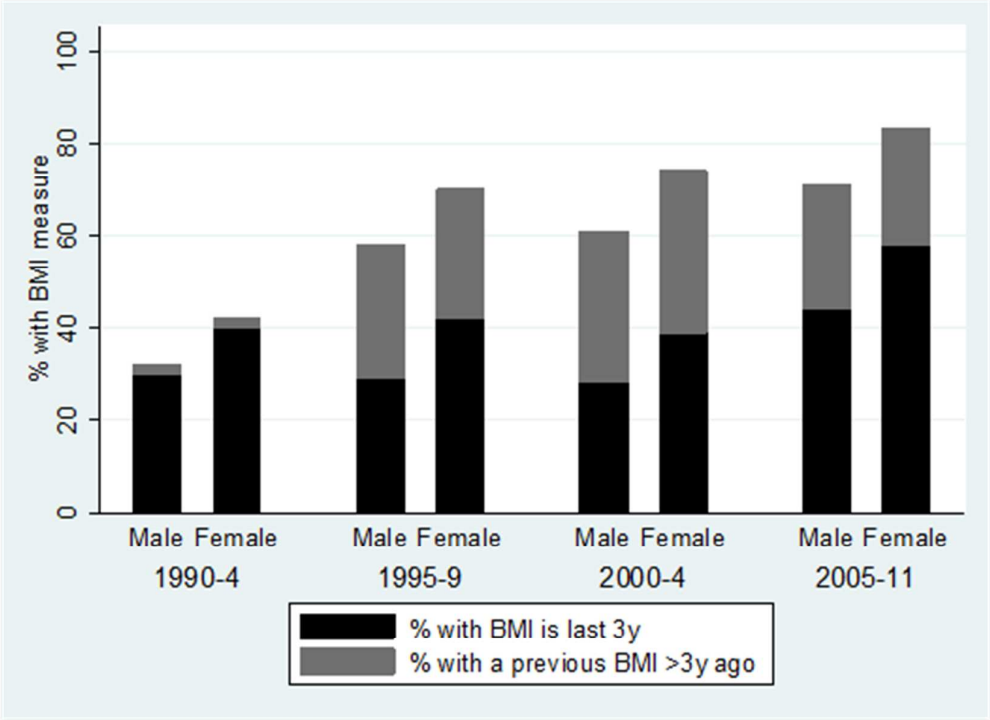
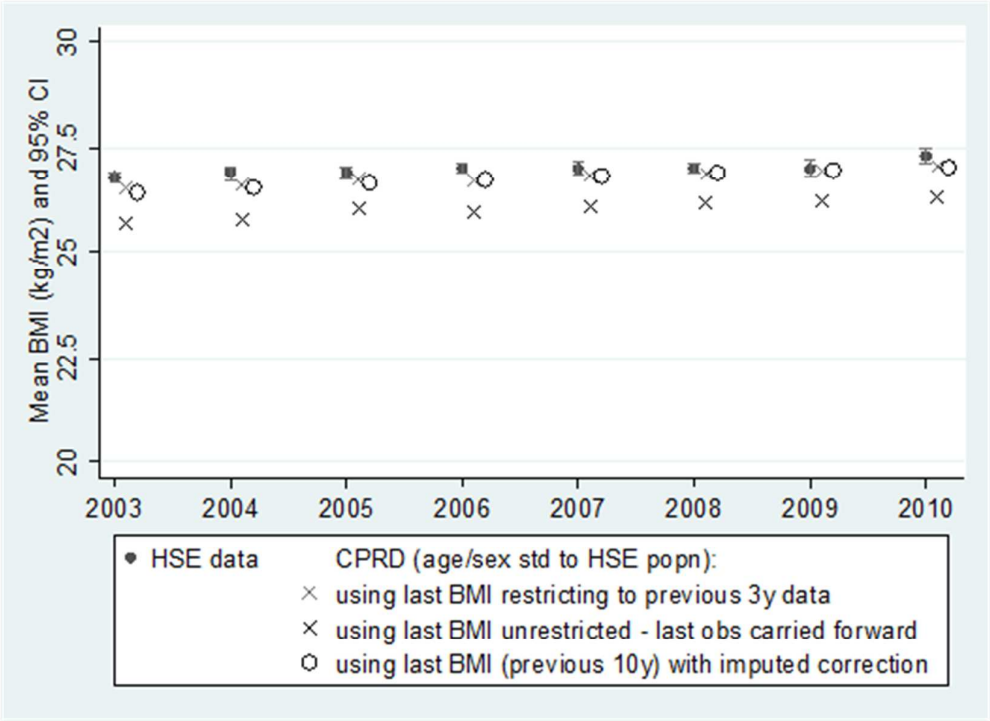


Figure 2: Completeness of BMI data in CPRD, by gender and calendar period



Note: Completeness data for each calendar period are based on all those under follow-up at mid-point of the period

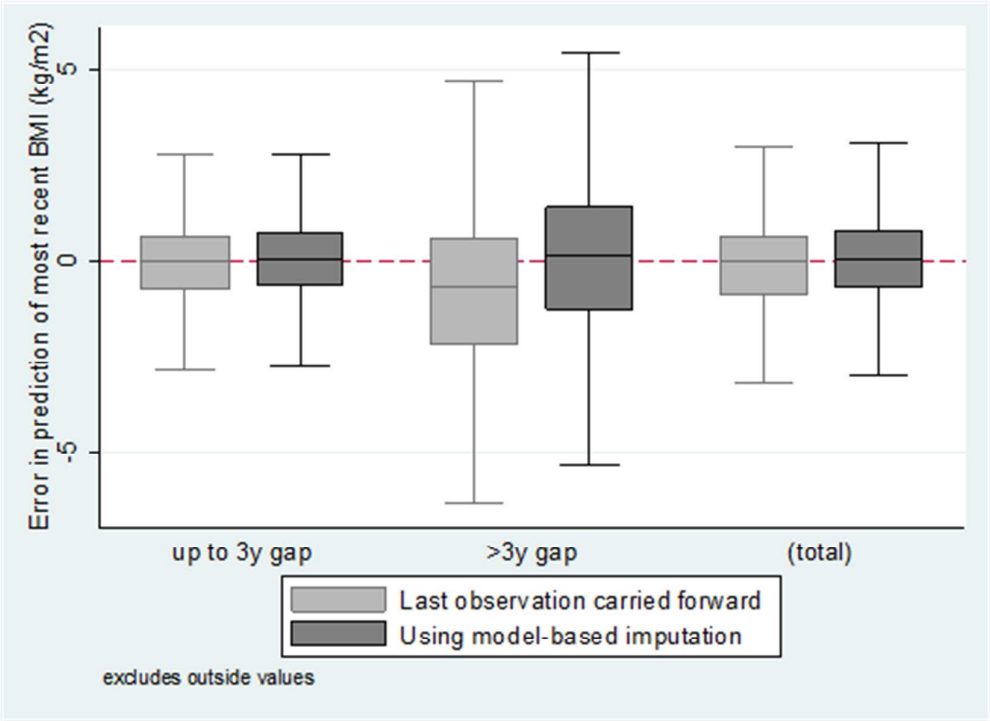
Figure 3: Mean BMI over calendar time comparing those with BMI recorded in CPRD (English practices) with the Health Survey for England 2010 data



Note: CPRD figures are age- and sex- standardised to the Health Survey for England study population

CPRD statistics are based on all patients registered at the mid-point of the calendar period and with a suitable previous BMI measure available (i.e. either any previous, or within the last 3 years)

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



Author contributions

I, Krishnan Bhaskaran, developed the analytical strategy for this paper, processed and analysed the data and wrote the paper.

I, Harriet Forbes, was involved in discussing the data processing and analysis of the data, as well as the writing of the paper.

I, Liam Smeeth, was involved in discussions of the analytical approach to this study and made comments on the analysis and the writing of the paper.

I, Ian Douglas, was involved in discussions of the analytical approach to this study and made comments on the analysis and the writing of the paper.

I, David Leon, was involved in discussions of the analytical approach to this study and made comments on the analysis and the writing of the paper.

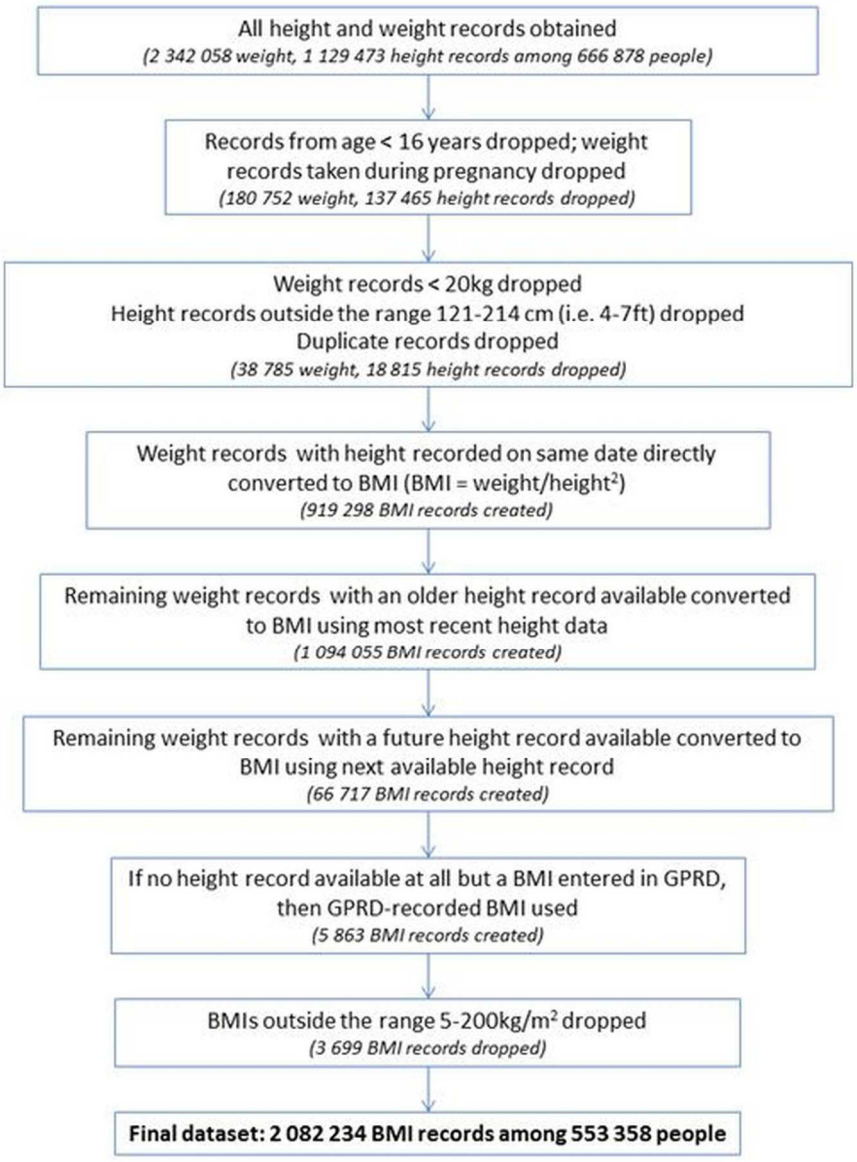


Figure 1: Initial data processing to generate BMI for analysis
141x187mm (300 x 300 DPI)

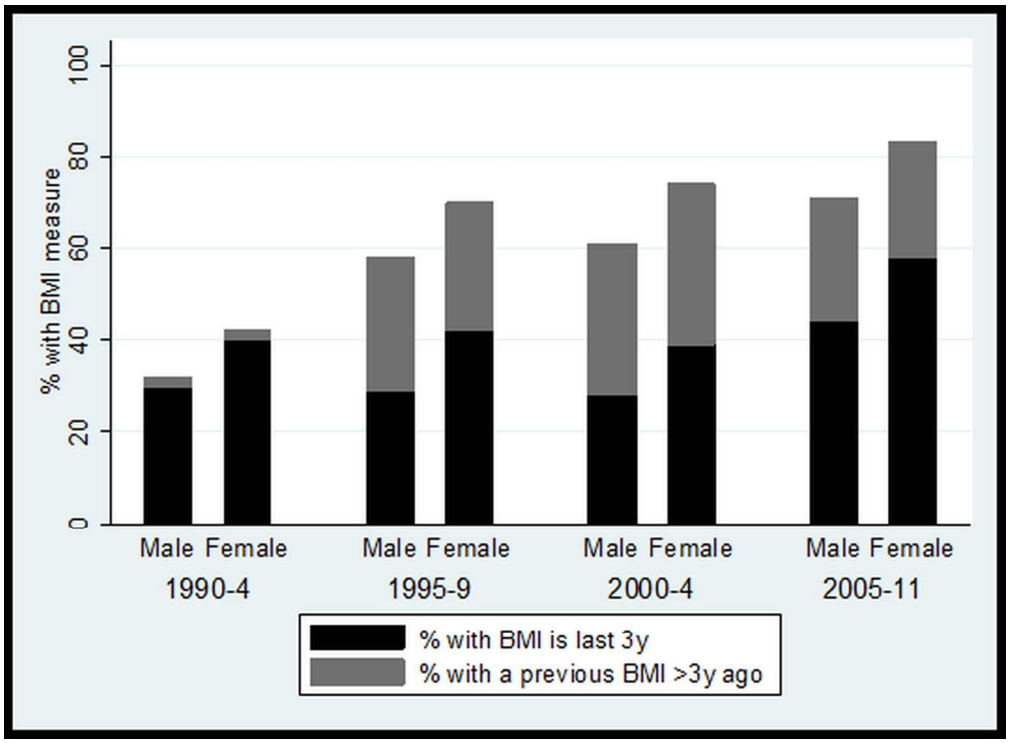


Figure 2: Completeness of BMI data in CPRD, by gender and calendar period
141x103mm (300 x 300 DPI)

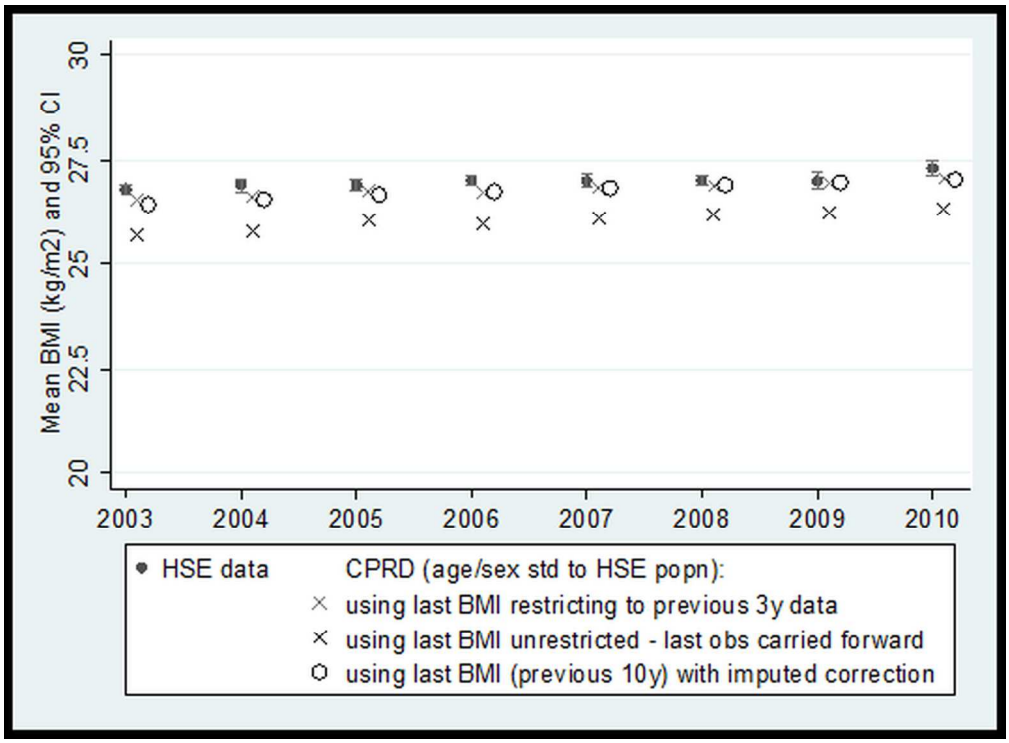


Figure 3: Mean BMI over calendar time comparing those with BMI recorded in CPRD (English practices) with the Health Survey for England 2010 data
141x103mm (300 x 300 DPI)

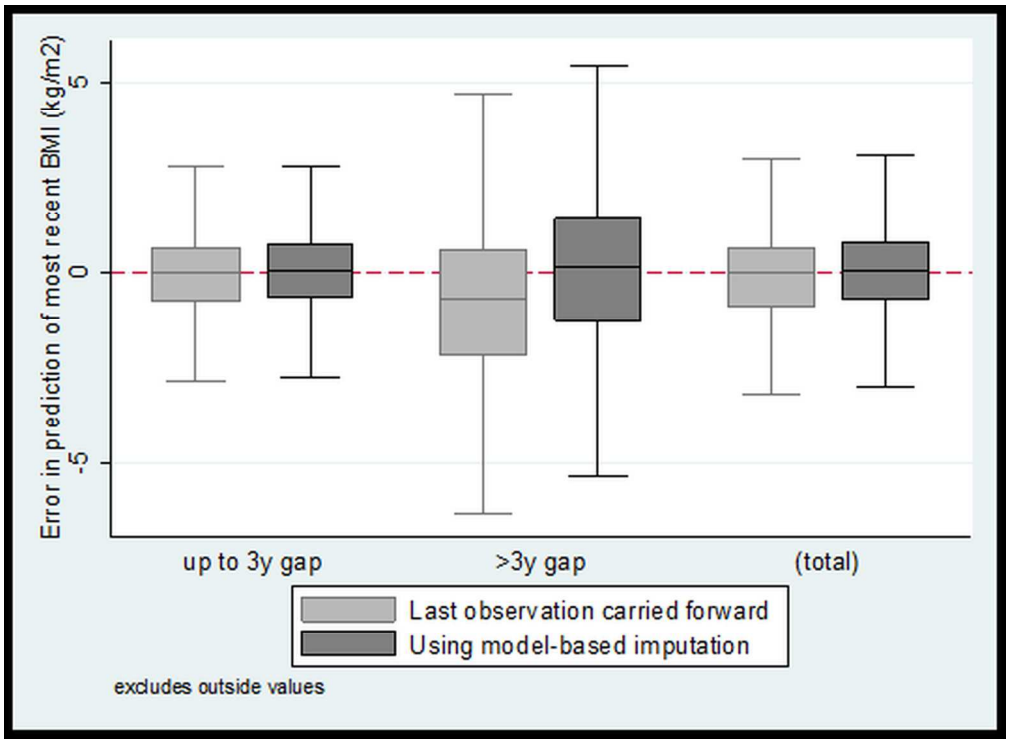


Figure 4: Error in prediction of most recent BMI from older BMI, comparing simple last observation carried forward with model-based imputation of up to date BMI – stratified by time gap between readings
141x103mm (300 x 300 DPI)

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

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

Faculty of Epidemiology & Population Health, London School of Hygiene and Tropical Medicine, Keppel Street, London, WC1E 7HT, UK

Corresponding Author: Krishnan Bhaskaran, London School of Hygiene and Tropical Medicine, Keppel Street, London, WC1E 7HT. Email: krishnan.bhaskaran@lshtm.ac.uk. Tel: + 44 (0) 20 7927 2268

Key words: CPRD, BMI, missing data, primary care databases, obesity.

Word Count: ~~3383~~3680

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 Clinical Practice Research Datalink (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 (2003-10) were age- 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 1990-94 to 77% in 2005-11, 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 0.75-1.1kg/m^2. Restricting to those with a recent (≤3 years) BMI resulted in mean BMI estimates closer to HSE (≤0.28kg/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 (≤0.37kg/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

Article focus:

- Body mass index (BMI) data are frequently used in epidemiological analyses of primary care databases such as the UK Clinical Practice Research Datalink (CPRD), however their completeness and representativeness have not previously been assessed in detail.
- The aim of this article is to provide information on the completeness of BMI in CPRD primary care data, on their representativeness, and on the implications for their practical use in research.

Key messages:

- We found that completeness of BMI recordings in the Clinical Practice Research Datalink increased from 37% in 1990-4 to 77% in 2005-11 and differed by age and sex.
- At specific calendar time points, the most recent BMI recorded for a large proportion of patients was over 3 years old and was unlikely to reflect current BMI.
- The optimal strategy for assigning BMI status is likely to depend on the specific study population and research context. We suggest one possible approach that uses a model-based imputation of current BMI to reduce BMI misclassification.

Strengths and limitations of this study:

- Results presented here are based on a large random sample from the CPRD, therefore we can confidently generalise the findings to the whole CPRD database, and to similar databases based on UK primary care records.
- To assess the representativeness of CPRD 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 both developed and developing countries[3 4] and is a growing concern to 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]. Because of 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 (i.e. recorded when the patient is attending for other reasons, or when 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)

The Clinical Practice Research Datalink (CPRD) is a clinical database comprising anonymised computerised medical records from general practitioners (GPs) in the United Kingdom. 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 a CPRD data ~~contains~~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.

Body mass 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). ~~Patient r~~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] The 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 8,420 in 2010),. ~~The methods are described in detail elsewhere.~~[19] ~~—~~Surveys were interviewer administered with interviewers measuring the weight and height of all participants. Data from 2003-10 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 Socio-economic 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 (1990-4, 1995-9, 2000-4, 2005-11). 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 mid-point 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 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 BMI category. We have not presented confidence intervals for these descriptive statistics because the sample size made sampling error negligible (for example, the standard errors for the proportions with complete BMI data in age and calendar year subgroups were all <0.5%).

Comparison of CPRD BMI data with Health Survey for England 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, HSE data were adjusted for non-response). CPRD mean BMI was based on patients registered and under up-to-standard follow-up at the mid-point of the calendar year. We produced two sets of CPRD mean BMI statistics: firstly we used last BMI observation carried forward (regardless of how long ago recorded); secondly we restricted to patients with a recent BMI available (up to 3 years before the mid-point of the calendar year). As above, confidence intervals are not presented because there was negligible sampling error (maximum standard error=0.02kg/m²). 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- and sex-standardised to the HSE population structure Proportions classified as obese (BMI≥30kg/m^2) over time 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 BMI recording (restricting to individuals with BMI records ≤ 10 years apart). We fitted a linear regression model with change in BMI as the outcome~~, and~~; the

main covariate predicting change in BMI was elapsed time, which was included as a 3 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 heteroskedasticity.[20]

Having specified a model for change in BMI over time, we first explored its performance among individuals with at least 2 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 from 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 included all individuals with a BMI record in the previous 10 years and used the model described above to impute current BMI at the mid-point 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 5 imputations) to account for uncertainty in the modelled changes over time.[21]

The study was approved by the London School of Hygiene and Tropical Medicine Ethics Committee.

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 2005-11(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 2005-11). 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 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 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 (Appendix Table A2). This is in line with 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, out of 75518 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 (7675/250430 = 3.0% in the most recent period).

Summary statistics using complete CPRD BMI data and comparison with Health Survey for England

We found that age- and sex-standardised mean BMI based on CPRD data was consistently and substantially lower (by up to 1.1kg/m^2) than in the HSE data (mean BMI in CPRD = 25.7kg/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 to 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 CPRD data underestimating the HSE statistics by only between 0.04 to 0.28kg/m^2 in individual calendar

years, and the CPRD estimate falling within the HSE confidence interval for 2 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- and sex-stratified data demonstrated similar patterns, except that in the eldest age group (75+ years), restricting to those with recent BMI measures did not bring the estimated BMI substantially closer to HSE figures (Appendix Figure A1).

We also compared the proportions classified as obese between CPRD and HSE (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 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 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 (Appendix Figure A3). We tested the predictive performance of our model by predicting the most recent BMI based on the previous one, among CPRD patients 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 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 = -0.70kg/m^2 [IQR -2.18 to +0.56] using last observation carried forward, compared with +0.11kg/m^2 [-1.29 to +1.40] using 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. Estimated mean BMIs were now in line with those based on only recent data in the earlier analysis, and were only between 0.04 and 0.37kg/m^2 lower than HSE statistics in individual calendar years (Figure 3, circles). Even with multiple imputation, confidence intervals remained extremely narrow (<0.07kg/m^2) due to the large sample size, so 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 to 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 1990-94 to 77% in 2005-11). 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 equivalent HSE estimates (by up to 1.1kg/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, mean BMI statistics were closer to 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 (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 2005-11). As in our study, the QRESEARCH report showed an increase in completeness over time, rising from 42% in 2000/01 to 58% in 2007. In a third UK primary care database, THIN (The Health Improvement Network), 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 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 prevalence~~'~~s of diseases in which it is important to monitor BMI.

Comparison of CPRD BMI data with Health Survey for England data

Mean BMI based on the CPRD population was lower in each calendar year than 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 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 ~~these~~ our modelled changes over time brought CPRD mean BMI statistics in line with 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 (for example, 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 recent 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 numbers 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 done 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 two-fold algorithm which imputes missing data from longitudinal variables at particular time points by using adjacent data points.[30] Ultimately, the ~~importance of these issues~~ pros and cons of various methods, and the optimal strategy

to use is likely to depend on the particular study and the characteristics of the study population.

Strengths and Limitations

Results presented here are based on a large random sample from the CPRD, therefore we can confidently generalise the findings to the whole CPRD database. Although we cannot assume these findings will relate to other routinely collected primary care databases in UK based on other IT systems (CPRD is based on practices using VISION), they are likely to be similar. This study did not look at BMI recordings among children as this would require a different strategy. Completeness among 16-24 year 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 ~~(sample size 14,836 in 2003 and 8,420 in 2010),~~ 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 focussed on the mean (and in the appendix, on the proportion classed as obese); these are the principal statistics published in 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.

Conflicts of interest
The authors declare no conflicts of interest.

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.

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. In: Organisation WH, ed. Geneva, Switzerland, 2009.
2. Flegal KM, Graubard BI, Williamson DF, et al. Excess deaths associated with underweight, overweight, and obesity. Jama-J Am Med Assoc 2005;**293**(15):1861-67
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**(9793):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**(9):1431-37 doi: Doi 10.1038/Ijo.2008.102[published Online First: Epub Date]|.
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 doi: 1471-2261-7-38 [pii]

10.1186/1471-2261-7-38[published Online First: Epub Date]|.
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 doi: 10.1136/bmj.c4444[published Online First: Epub Date]|.
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. British Medical Journal 2009;**339**:b4731 doi: Artn B4731

Doi 10.1136/Bmj.B4731[published Online First: Epub Date]|.
10. Douglas I, Smeeth L, Irvine D. The use of antidepressants and the risk of haemorrhagic stroke: a nested case control study. Brit J Clin Pharmaco 2011;**71**(1):116-20 doi: DOI 10.1111/j.1365-2125.2010.03797.x[published Online First: Epub Date]|.
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 Psychiat 2009;**166**(5):591-98 doi: DOI 10.1176/appi.ajp.2008.08071065[published Online First: Epub Date]|.
12. Lawrenson R, Todd JC, Leydon GM, et al. Validation of the diagnosis of venous thromboembolism in general practice database studies. Brit J Clin Pharmaco 2000;**49**(6):591-96
13. Jick H, Zornberg GL, Jick SS, et al. Statins and the risk of dementia. Lancet 2000;**356**(9242):1627-31
14. van Staa TP, Wegman S, de Vries F, et al. Use of statins and risk of fractures. Jama-J Am Med Assoc 2001;**285**(14):1850-55
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, 2007:341-48.
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**(4):588-96 doi: 10.1111/j.1365-2125.2010.03733.x[published Online First: Epub Date]|.
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, New Jersey: Prentice Hall., 1997.
21. Rubin DB. *Multiple Imputation for Nonresponse in Surveys*: J. Wiley & Sons, New York. , 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**(6):618-26 doi: 10.1002/pds.1934[published Online First: Epub Date]|.
25. Rowlands S, Moser K. Consultation rates from the General Practice Research Database. Brit J Gen Pract 2002;**52**(481):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. Journal of Hypertension 2007;**25**(6):1215-23
28. Silverwood RJ, Pierce M, Thomas C, et al. Overweight across adult life and kidney function at age 60-4 years: the 1946 British birth cohort study. (awaiting publication) 2012
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**(14):1315-25 doi: 10.1056/NEJMoa1006992[published Online First: Epub Date]|.
30. Welch C, Bartlett J, Petersen I. Application of multiple imputation using the two-fold fully conditional specification algorithm in longitudinal clinical data. (In Press). Stata Journal 2013
31. Mindell J, Biddulph JP, Hirani V, et al. Cohort Profile: The Health Survey for England. International Journal of Epidemiology 2012:1-9 doi: 10.1093/ije/dyr199[published Online First: Epub Date]|.

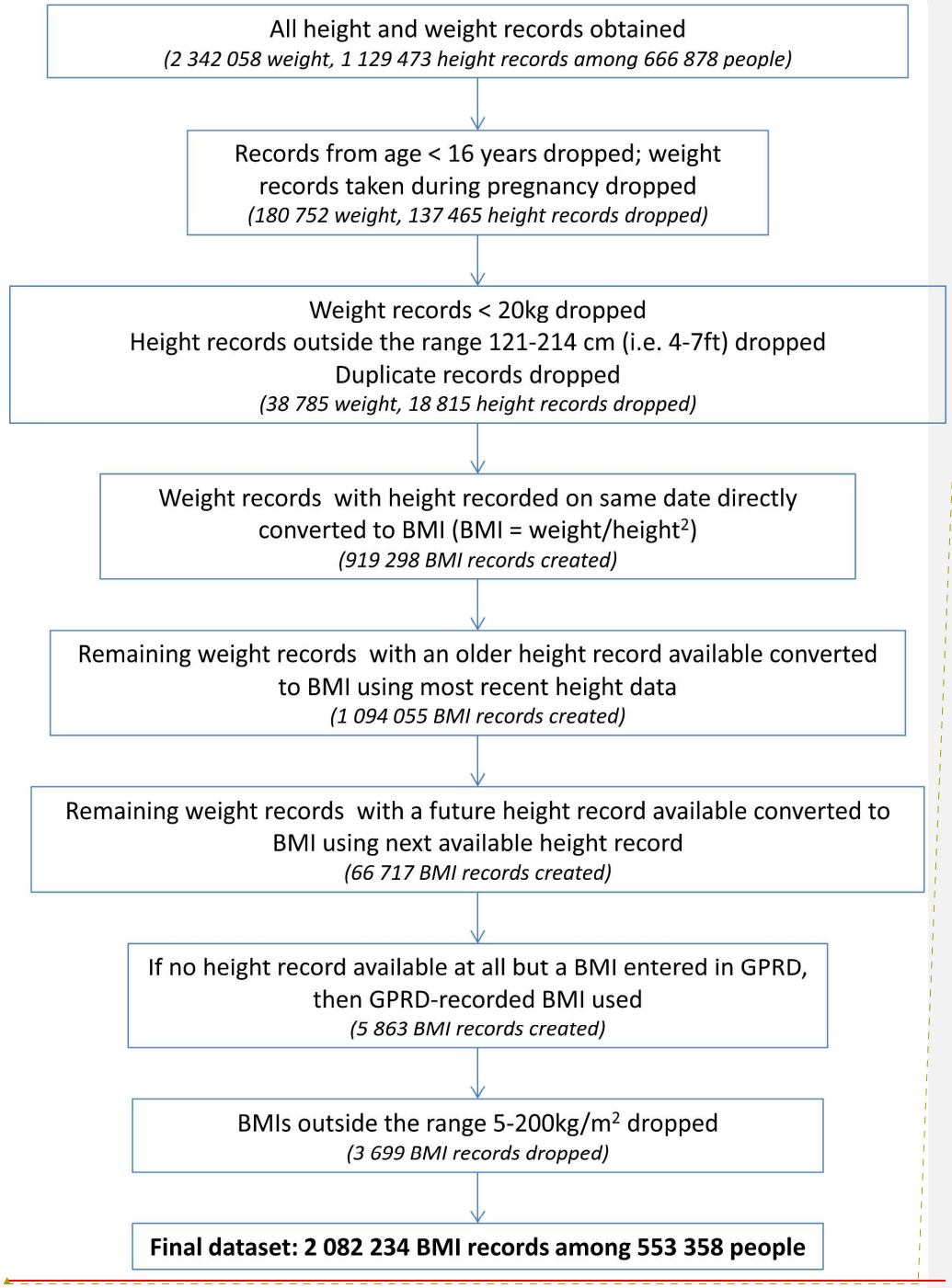
Table 1: Completeness of BMI data in the CPRD, by age and calendar period

Age group (yrs)	1990-4	1995-9	2000-4	2005-2011
16-24[a]				
N registered	11423	17501	34452	42546
BMI in previous 3y (%)	26	28	25	32
BMI any previous (%)	26	37	30	37
25-34				
N registered	17477	29923	48659	50413
BMI in previous 3y (%)	37	39	36	49
BMI any previous (%)	38	66	67	72
35-44				
N registered	15953	28838	55991	61014
BMI in previous 3y (%)	36	36	31	46
BMI any previous (%)	39	67	71	80
45-54				
N registered	14507	27765	48093	55564
BMI in previous 3y (%)	39	37	32	50
BMI any previous (%)	42	70	73	84
55-64				
N registered	11680	20843	42258	49380
BMI in previous 3y (%)	42	40	37	57
BMI any previous (%)	44	74	77	87
65-74				
N registered	10678	17605	30997	34508
BMI in previous 3y (%)	36	37	40	67
BMI any previous (%)	38	71	79	91
75+				
N registered	8637	16005	29384	32523
BMI in previous 3y (%)	28	32	37	64
BMI any previous (%)	28	56	69	87
Total				
N registered	90355	158480	289834	325948
BMI in previous 3y (%)	35	36	34	51
BMI any previous (%)	37	64	67	77

N registered is all those under follow-up at mid-point of the period

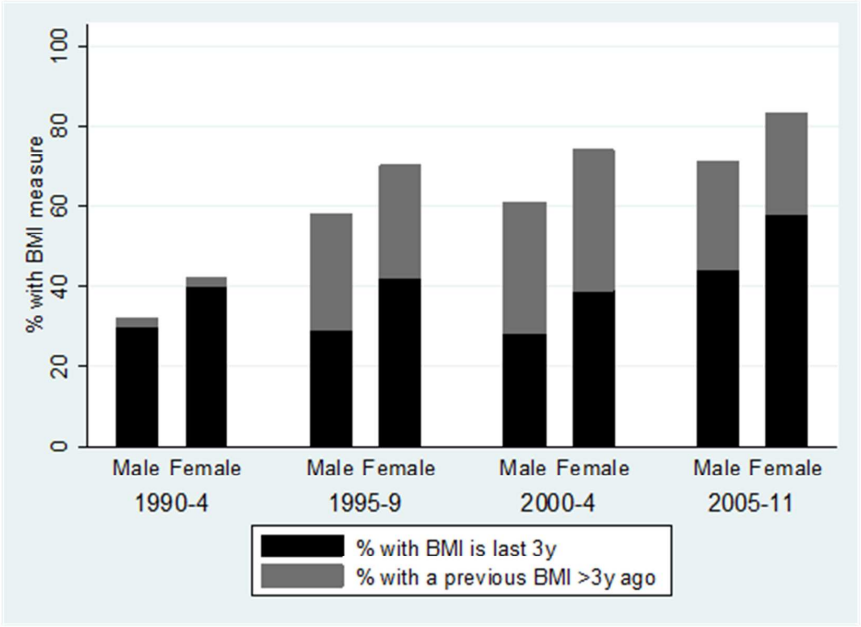
[a]Note, BMI measurements from age <16 years were not counted in this analysis, hence completeness in the 16-24 age group may be artificially low

Figure 1: Initial data processing to generate BMI for analysis



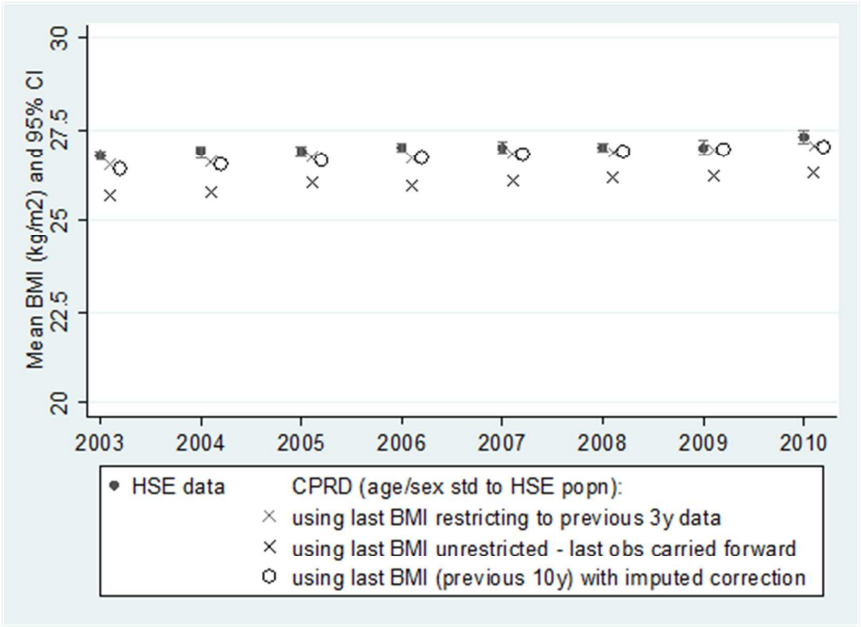
Formatted: Font: 12 pt

Figure 2: Completeness of BMI data in CPRD, by gender and calendar period



Note: Completeness data for each calendar period are based on all those under follow-up at mid-point of the period

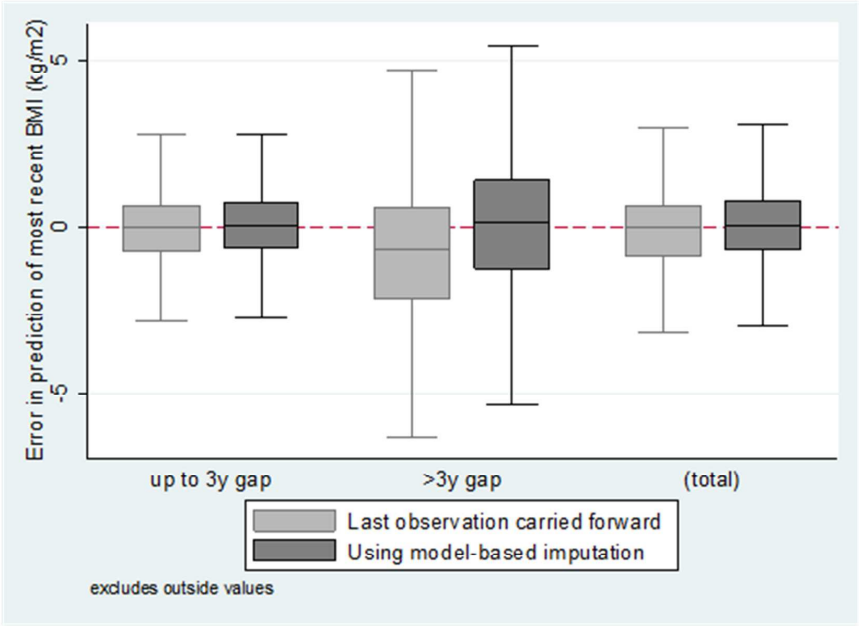
Figure 3: Mean BMI over calendar time comparing those with BMI recorded in CPRD (English practices) with the Health Survey for England 2010 data



Note: CPRD figures are age- and sex- standardised to the Health Survey for England study population

CPRD statistics are based on all patients registered at the mid-point of the calendar period and with a suitable previous BMI measure available (i.e. either any previous, or within the last 3 years)

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



Author contributions

I, Krishnan Bhaskaran, developed the analytical strategy for this paper, processed and analysed the data and wrote the paper.

I, Harriet Forbes, was involved in discussing the data processing and analysis of the data, as well as the writing of the paper.

I, Liam Smeeth, was involved in discussions of the analytical approach to this study and made comments on the analysis and the writing of the paper.

I, Ian Douglas, was involved in discussions of the analytical approach to this study and made comments on the analysis and the writing of the paper.

I, David Leon, was involved in discussions of the analytical approach to this study and made comments on the analysis and the writing of the paper.

STROBE Statement—Checklist of items that should be included in reports of ***cohort studies***

NOTE Page numbers refer to revised manuscript, tracked changes version

	Item No	Recommendation
Title and abstract	1	(*a*) Indicate the study's design with a commonly used term in the title or the abstract n/a (we did not think there was an appropriate design keyword/term to describe this study as it is not a standard "exposure/outcome" study but is rather providing data quality information on a common exposure/covariate)
		(*b*) Provide in the abstract an informative and balanced summary of what was done and what was found P2
Introduction		
Background/rationale	2	Explain the scientific background and rationale for the investigation being reported P4-5
Objectives	3	State specific objectives, including any prespecified hypotheses P4-5
Methods		
Study design	4	Present key elements of study design early in the paper P6-7
Setting	5	Describe the setting, locations, and relevant dates, including periods of recruitment, exposure, follow-up, and data collection P5-6
Participants	6	(*a*) Give the eligibility criteria, and the sources and methods of selection of participants. Describe methods of follow-up P5-6
		(*b*) For matched studies, give matching criteria and number of exposed and unexposed
Variables	7	Clearly define all outcomes, exposures, predictors, potential confounders, and effect modifiers. Give diagnostic criteria, if applicable P5-7
Data sources/ measurement	8*	For each variable of interest, give sources of data and details of methods of assessment (measurement). Describe comparability of assessment methods if there is more than one group P5-6
Bias	9	Describe any efforts to address potential sources of bias P6-8
Study size	10	Explain how the study size was arrived at P5
Quantitative variables	11	Explain how quantitative variables were handled in the analyses. If applicable, describe which groupings were chosen and why P6-7
Statistical methods	12	(*a*) Describe all statistical methods, including those used to control for confounding P6-7
		(*b*) Describe any methods used to examine subgroups and interactions P6-7
		(*c*) Explain how missing data were addressed

		P7
		(*d*) If applicable, explain how loss to follow-up was addressed n/a
		(*e*) Describe any sensitivity analyses n/a
Results		
Participants	13*	(a) Report numbers of individuals at each stage of study—eg numbers potentially eligible, examined for eligibility, confirmed eligible, included in the study, completing follow-up, and analysed
		(b) Give reasons for non-participation at each stage
		(c) Consider use of a flow diagram FIG 1
Descriptive data	14*	(a) Give characteristics of study participants (eg demographic, clinical, social) and information on exposures and potential confounders
		(b) Indicate number of participants with missing data for each variable of interest
		(c) Summarise follow-up time (eg, average and total amount) P8-9 and FIG 2
Outcome data	15*	Report numbers of outcome events or summary measures over time n/a (no specific outcome)
Main results	16	(*a*) Give unadjusted estimates and, if applicable, confounder-adjusted estimates and their precision (eg, 95% confidence interval). Make clear which confounders were adjusted for and why they were included
		(*b*) Report category boundaries when continuous variables were categorized
		(*c*) If relevant, consider translating estimates of relative risk into absolute risk for a meaningful time period n/a (not an "exposure/outcome" study)
Other analyses	17	Report other analyses done—eg analyses of subgroups and interactions, and sensitivity analyses P9-11
Discussion		
Key results	18	Summarise key results with reference to study objectives P11-12
Limitations	19	Discuss limitations of the study, taking into account sources of potential bias or imprecision. Discuss both direction and magnitude of any potential bias P15-16
Interpretation	20	Give a cautious overall interpretation of results considering objectives, limitations, multiplicity of analyses, results from similar studies, and other relevant evidence P16
Generalisability	21	Discuss the generalisability (external validity) of the study results P15
Other information		
Funding	22	Give the source of funding and the role of the funders for the present study and, if

	applicable, for the original study on which the present article is based
	P17

*Give information separately for exposed and unexposed groups.

Note: An Explanation and Elaboration article discusses each checklist item and gives methodological background and published examples of transparent reporting. The STROBE checklist is best used in conjunction with this article (freely available on the Web sites of PLoS Medicine at http://www.plosmedicine.org/, Annals of Internal Medicine at http://www.annals.org/, and Epidemiology at http://www.epidem.com/). Information on the STROBE Initiative is available at http://www.strobe-statement.org.