BMJ Open
Evaluation of routinely collected records for dementia outcomes in UK: a prospective cohort study

Shabina Hayat 1, Robert Luben,2,3 Kay-Tee Khaw,2 Nicholas Wareham,2 Carol Brayne1

ABSTRACT
Objectives To evaluate the characteristics of individuals recorded as having a dementia diagnosis in different routinely collected records and to examine the extent of overlap of dementia coding across data sources. Also, to present comparisons of secondary and primary care records providing value for researchers using routinely collected records for dementia outcome capture.

Study design A prospective cohort study.

Setting and participants A cohort of 25 639 men and women in Norfolk, aged 40–79 years at recruitment (1993–1997) followed until 2018 linked to routinely collected to identify dementia cases. Data sources include mortality from death certification and National Health Service (NHS) hospital or secondary care records. Primary care records for a subset of the cohort were also reviewed.

Primary outcome measure Diagnosis of dementia (any-cause).

Results Over 2000 participants (n=2635 individuals) were found to have a dementia diagnosis recorded in one or more of the data sources examined. Limited concordance was observed across the secondary care data sources. We also observed discrepancies with primary care records for the subset and report on potential linkage-related selection bias.

Conclusions Use of different types of record linkage from varying parts of the UK’s health system reveals differences in recorded dementia diagnosis, indicating that dementia can be identified to varying extents in different parts of the NHS system. However, there is considerable variation, and limited overlap in those identified. We present potential selection biases that might occur depending on whether cause of death, or primary and secondary care data sources are used. With the expansion of using routinely collected health data, researchers must be aware of these potential biases and inaccuracies, reporting carefully on the likely extent of limitations and challenges of the data sources they use.

INTRODUCTION
 Routinely collected health records are powerful tools that have been increasingly used for research in the last decade.1 Epidemiological studies depend on accurate, reliable and relevant information from data sources. However, there is substantial heterogeneity in the methods for case-ascertainment in dementia literature,2 with studies applying different criteria, with no ‘gold standard’.3 Accuracy of medical records, of whatever type, tends to be assumed in the literature. Compared with population based surveys, dementia is still known to be under-reported in primary healthcare records, with individuals having manifest signs of dementia but no record of a dementia diagnosis, although this ‘gap’ is being closed more recently.4 Dementia ascertainment is complex in primary and secondary care settings5 6 as well on death certification.7 There are many influences on whether dementia is likely to be recorded in these settings. Variations in dementia reporting have been noted across regions in UK.8 9 Data sources are prone to inconsistency, misclassification and are influenced by change in dementia practice and policy,5 such as those summarised in box 1.

In the UK, the National Health Service (NHS) holds data on all hospital admissions by condition, allowing dementia ascertainment from routine data sources. To date, mortality and Hospital Episode Statistics (HES) data have been the main sources of health records for epidemiological studies.3 10 11 These data are coded using the International Classification of Diseases (ICD) which are mainly diagnostic codes. Using these sources alone seriously underestimates the number of dementia cases, as most dementia patients will not require hospitalisation for their dementia nor will they have dementia as a cause of death recorded on their death certificate.12

General practitioners (GP) maintain primary care records, which include information, both for diagnoses and also for administrative purposes, including details of specialist referrals.
such as memory clinics. Diagnosis of dementia is usually initiated in a primary care setting, based on patient symptoms or caregivers’ concerns. GP records, in theory, should be a more complete source for dementia case ascertainment. The coding system used in primary care uses Read Codes, is separate from the ICD system used in secondary care and death registration. Codes for dementia are numerous, complex and prone to coding error and misclassification. Another data source for dementia, which seems to have had confusion of dementia case ascertainment. The impact of incomplete or inaccurate recording of dementia can affect the generalisability of findings. Furthermore, databases are influenced by the point data are collected in the patient care pathway and may not be representative of the full population of people living with dementia, thus affecting generalisability of findings.

The objective of this study was to describe the procedures for a better understanding of the characteristics of individuals recorded as having a dementia diagnosis in different health care records and examine the extent of overlap of dementia coding across data sources. We present novel insight into the MHDS and compare secondary and primary care records providing value for others relying on these types of data sources for outcome capture.

**Box 1 Changes in policy for public and practitioners which may have increased awareness of dementia and recording of cases in health records**

2006: Introduction of dementia to Quality and Outcomes Framework (QOF) registers. QOF is an incentivising primary care scheme that rewards general practitioner practices for the provision of quality care and helps to standardise improvements in the delivery of primary medical services. Participation in QOF is voluntary.

2009: The 2009 National Dementia Strategy was introduced to increase the public and professional awareness and understanding of dementia. Also, the dementia prevalence indicator to the QOF was introduced in primary care.

2011: Changes to coding practice. For mortality coding, deaths previously coded with an underlying cause of unspecified cerebrovascular disease were reclassified as vascular dementia.

2012: The Dementia Challenge was launched in March 2012 by Prime Minister, David Cameron. This programme’s, which superseded the national strategy from 2009, focus was on three main areas: bringing about improvements in health and care, creating dementia friendly communities and improving research.

2013: The Dementia Commissioning for Quality and Innovation was introduced to encourage practitioners to identify those who potentially have dementia in a secondary care setting, most particularly applied to those over 75 years with unplanned hospital admissions.

2014: Mortality coding revised—dementia coded as the underlying cause of death.

2015: Prime Minister launched his Challenge on Dementia 2020, which set out to build on the achievements of the Prime Minister’s Challenge on Dementia 2012–2015.

**METHODS**

**Study population and data collection**

Participants were recruited to the European Prospective Investigation of Cancer (EPIC) in Norfolk. Details of this study have been previously described. Briefly, 50,445 men and women (40–79 years) residents of Norfolk, England, were recruited to the study via their general practices. Of these, 25,639 attended the baseline health examination between 1993 and 1997. Subsequent follow-ups have involved self-report of health and lifestyle postal questionnaires and further health examinations. Virtually complete follow-up for cohort for mortality and hospital admissions in EPIC-Norfolk has been established via linkage to routinely collected NHS databases in England (HES) and for mortality data for all participants using their unique NHS number and date of birth.

**Dementia ascertainment and diagnostic codes in hospital and death records and mental healthcare data**

Data linkage to the EPIC-Norfolk cohort for NHS and mortality data was carried out by NHS Digital, a statutory body in England, permitted to receive identifiable patient data for linkage. The linked hospital records contain coded diagnostic information for all inpatient and day-case admissions. We also obtained national MHDS, which contain record-level data about individuals in contact with mental health services including memory clinics.

Participants with incident dementia were defined as those free of dementia at the time of enrolment to the study and then identified with a dementia diagnosis recorded in medical records subsequently. For a subset, we compared the NHS Digital data sets of secondary care and mortality to GP records. GP records for individuals with a dementia diagnosis in secondary care but missing from data provided by practices were further reviewed by researchers in the third and final confirmation phase for any indication of a dementia code or mention in free text. Further details of the protocol and dementia codes are provided in the online supplemental information and table S1. QOF for dementia coding in primary care financially incentivises each recorded dementia diagnosis. Figure 1 is a diagrammatic representation of selection of participants and record linkage in the cohort.

**Ethics**

EPIC-Norfolk has the permissions (Section 251 and where possible, explicit signed consent given by participants attending health examinations) to follow-up through medical record linkage. Procedures for this study have been approved by NHS Research Ethics Committee (NRES REC Ref: 98CN01). The Regulatory approvals from the Confidentiality Advisory Group (CAG) can be found at:

Section 251 Application number 059;  https://www.hra.nhs.uk/planning-and-improving-research/application-summaries/confidentiality-advisory-group-registers/.

For the GP dementia case ascertainment and validation substudy—we also obtained approval from Norfolk &
RESULTS

There were 2635 cases of dementia identified from the cohort of 25 639 individuals at the censor date of 31 December 2018 after 25.8 years of follow-up. The youngest age of entry to the study at baseline was just below 40 and the oldest age of the participant at the censor date was 101 years. Out of the 2298 individuals with data on age of diagnosis, the minimum age of diagnosis was 54 years and maximum was 99 years. Figure 2 shows the relationship between age and dementia diagnosis from mid to later life. Table 1 shows the sex-specific and age-specific cumulative incidence of dementia and deaths in the cohort. Increasing age was associated with increasing rates of dementia and death. This table reflects

Figure 2 Scatter plot of age of diagnosis by age at baseline in the EPIC-Norfolk cohort. EPIC, European Prospective Investigation of Cancer.
the higher mortality in men which results in higher absolute numbers and of dementia cases in women.

On comparing characteristics across the different data sources, there were no major differences in terms of sex and sociodemographic profiles (table 2). Those with dementia recorded from the mortality data set were slightly older and in MHDS were younger. Most recorded diagnoses of dementia were found in HES, followed by mortality records, with the least number identified from the MHDS. The distribution of dementia diagnoses across the three NHS Digital data sources is shown in figure 3. This figure shows very little overlap. Where there was concordance in cases for the MHDS and HES data sources (n=448), 71% (319) had a date of diagnosis in the MHDS before the date of diagnosis in HES (median (IQR) 0.7 years (0.3, 1.7 years) with 26% (116) with an earlier date in HES and median (IQR) 0.6 years (0.2, 1.9 years). Only 3% (13) had the same date recorded in both.

Associations for sociodemographic factors and the presence of a recorded diagnosis of dementia were similar across the three data sources (table 3). Age was a stronger predictor in the mortality data and weakest in the mental health data. Having qualifications was associated with a lower risk of future recorded diagnosis of dementia, this was observed for mortality and HES data but not for the MHDS.

The MHDS consisted of mainly administrative data such as mental health reviews, care programmes and pathways that include contacts with mental healthcare professionals (both in hospitals and in outpatient memory clinics and the community) as well as diagnostics and treatment codes. The service-level breakdown of the mental health data was not applicable here, as there was little additional diagnostic information. The latest release in 2017–2018 appeared to be the most complex, covering mental healthcare more comprehensively and containing diagnostic ICD10 codes that had been limited in the previous years.

For the substudy involving validation with GP records, out of the 26 practices that were contacted, 14 agreed to participate, 6 declined and 6 did not respond. Of the 14 that agreed, 8 practices completed the questionnaire and 9 provided data. There were no criteria for selecting a practice for this validation, other than they were practices that were collaborating with the study. There was a mix of rural and city practices, classified as urban or town and fringe areas, although the majority of the practices ultimately submitting data were city-based. Different types of staff carried out the linkage with their records, ranging from practice managers, IT managers, research nurses to a healthcare assistant.

In this subpopulation, 4.4% (209 of the 4668 participants) were dementia cases identified from HES, mental health data, mortality and the GP records compared with 10.1% (2635) dementia cases out of 25 635 participants in the rest of the cohort. The individuals of the substudy registered with the responding practices were younger, more likely to be women, with higher education and social class when compared with the rest of the cohort (online supplemental table S2). Of the 209 respondents with a recorded diagnosis of dementia, 57

---

**Table 1** Age-specific and sex-specific proportions of dementia and death in EPIC-Norfolk from 1996 until 31 December 2018 using all three secondary care data sources provided by NHS digital for each age group by gender at baseline

<table>
<thead>
<tr>
<th>Age band at baseline</th>
<th>Median age</th>
<th>Freq (N)</th>
<th>% with dementia recorded at any time in follow-up (N)</th>
<th>% Died (N)</th>
<th>P value</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Men (n=11 607)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>≤59 years</td>
<td>51.6</td>
<td>5915</td>
<td>3.2 (190)</td>
<td>17.4 (1028)</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>60–64 years</td>
<td>62.5</td>
<td>1848</td>
<td>12.3 (228)</td>
<td>46.7 (863)</td>
<td></td>
</tr>
<tr>
<td>65–69 years</td>
<td>67.5</td>
<td>1890</td>
<td>15.6 (294)</td>
<td>72.0 (1361)</td>
<td></td>
</tr>
<tr>
<td>70–74 years</td>
<td>72.5</td>
<td>1579</td>
<td>17.3 (273)</td>
<td>89.9 (1420)</td>
<td></td>
</tr>
<tr>
<td>&gt;75 years</td>
<td>75.7</td>
<td>375</td>
<td>14.9 (56)</td>
<td>95.7 (359)</td>
<td></td>
</tr>
<tr>
<td><strong>Women (n=14 032)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>≤59 years</td>
<td>51.3</td>
<td>7656</td>
<td>2.9 (222)</td>
<td>11.9 (909)</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>60–64 years</td>
<td>62.6</td>
<td>2118</td>
<td>14.5 (308)</td>
<td>36.9 (782)</td>
<td></td>
</tr>
<tr>
<td>65–69 years</td>
<td>67.4</td>
<td>2103</td>
<td>22.6 (475)</td>
<td>55.4 (1166)</td>
<td></td>
</tr>
<tr>
<td>70–74 years</td>
<td>72.4</td>
<td>1766</td>
<td>27.0 (477)</td>
<td>80.4 (1419)</td>
<td></td>
</tr>
<tr>
<td>&gt;75 years</td>
<td>75.8</td>
<td>389</td>
<td>28.8 (112)</td>
<td>90.7 (353)</td>
<td></td>
</tr>
<tr>
<td><strong>All (n=25 639)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>≤59 years</td>
<td>51.5</td>
<td>13 571</td>
<td>3.0 (412)</td>
<td>14.3 (1937)</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>60–64 years</td>
<td>62.6</td>
<td>3966</td>
<td>13.5 (536)</td>
<td>41.5 (1645)</td>
<td></td>
</tr>
<tr>
<td>65–69 years</td>
<td>67.5</td>
<td>3993</td>
<td>19.3 (769)</td>
<td>63.3 (2527)</td>
<td></td>
</tr>
<tr>
<td>70–74 years</td>
<td>72.5</td>
<td>3345</td>
<td>22.8 (750)</td>
<td>84.7 (2839)</td>
<td></td>
</tr>
<tr>
<td>&gt;75 years</td>
<td>75.7</td>
<td>764</td>
<td>22.0 (168)</td>
<td>93.2 (712)</td>
<td></td>
</tr>
</tbody>
</table>

P values by χ² for proportion of individuals with dementia in each age category. Results shown for men and women separately and combined (all). EPIC, European Prospective Investigation of Cancer; NHS, National Health Service.

<http://bmjopen.bmj.com/>
Table 2  Comparison of characteristics of ‘definite dementia’ cases identified from the three data sources separately and combined

<table>
<thead>
<tr>
<th></th>
<th>All three data sources</th>
<th>HES/MHDS/mortality</th>
<th>HES</th>
<th>MHDS</th>
<th>Mortality</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>n=2635</td>
<td>n=2157</td>
<td>n=727</td>
<td>n=1276</td>
<td>P value</td>
</tr>
<tr>
<td><strong>Sociodemographic</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean (SD)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Age at baseline</td>
<td>66.7 (6.5)</td>
<td>66.9 (6.3)</td>
<td>64.1 (6.5)</td>
<td>67.9 (5.9)</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Age at diagnosis</td>
<td>83.8 (6.5)</td>
<td>83.7 (6.5)</td>
<td>82.7 (6.8)</td>
<td>84.6 (6.4)</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Sex, % women (n)</td>
<td>60.5 (1594)</td>
<td>59.8 (1290)</td>
<td>59.7 (434)</td>
<td>60.5 (772)</td>
<td>0.9</td>
</tr>
<tr>
<td><strong>Education, % (n)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>No qualifications</td>
<td>48.4 (1274)</td>
<td>49.1 (1057)</td>
<td>45.3 (329)</td>
<td>50.4 (641)</td>
<td>0.5</td>
</tr>
<tr>
<td>O/A level standard</td>
<td>42.6 (1120)</td>
<td>41.9 (903)</td>
<td>45.3 (329)</td>
<td>41.4 (527)</td>
<td></td>
</tr>
<tr>
<td>Graduate level</td>
<td>9.0 (238)</td>
<td>9.0 (194)</td>
<td>9.4 (68)</td>
<td>8.2 (105)</td>
<td></td>
</tr>
<tr>
<td><strong>Social class, % (n)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Professional</td>
<td>6.3 (162)</td>
<td>6.3 (132)</td>
<td>5.6 (42)</td>
<td>6.0 (74)</td>
<td>0.7</td>
</tr>
<tr>
<td>Managerial</td>
<td>33.9 (866)</td>
<td>33.6 (702)</td>
<td>32.0 (240)</td>
<td>34.0 (421)</td>
<td></td>
</tr>
<tr>
<td>Skilled non-manual</td>
<td>20.5 (523)</td>
<td>20.5 (428)</td>
<td>21.1 (158)</td>
<td>21.7 (268)</td>
<td></td>
</tr>
<tr>
<td>Skilled manual</td>
<td>21.3 (544)</td>
<td>21.2 (442)</td>
<td>21.4 (160)</td>
<td>20.5 (253)</td>
<td></td>
</tr>
<tr>
<td>Semi-skilled</td>
<td>14.2 (362)</td>
<td>14.1 (295)</td>
<td>17.2 (129)</td>
<td>13.6 (168)</td>
<td></td>
</tr>
<tr>
<td>Non-skilled</td>
<td>3.8 (98)</td>
<td>4.2 (88)</td>
<td>2.7 (20)</td>
<td>4.3 (53)</td>
<td></td>
</tr>
</tbody>
</table>

HES, Hospital Episode Statistics; MHDS, mental health data set; N, number; O/A, ordinary or advanced level.

were found in both secondary and primary care data and 87 were in primary care records only. In this small study, almost all the practices were using Read codes and all reported their dementia cases to be confirmed through secondary referral to memory clinics. Five out of the eight practices reported that some dementia cases were not previously identified and so were missing on their QOF registers. The summarised responses to the questionnaire from GPs are given in online supplemental information table S3. Participants who were prescribed one of the four dementia drugs (n=57), also had a Read code of a definite dementia diagnosis. There were no cases identified from drugs alone.

There were 65 participants who had a definite dementia diagnosis in the secondary care records, but not in their GP records. For most of these (51 cases), the reason for absence from GP systems was because the participant had died, and the dementia diagnoses had come from mortality records. For the remaining (n=14), 3 participants had Mild Cognitive Impairment or probable dementia and for 10 participants, we were unable to confirm diagnosis. This could be because they had died recently, or had moved to another practice, in which case their records would no longer be available to the GP. One patient had no indication of any dementia or dementia related condition in their record. Of the 57 cases that were in both the secondary and primary data sources, 30 cases were obtained from the initial GP data extract, with 27 missed by the practices. These cases were ascertained with further review of individual records by the researchers and would have been missed if the extra confirmation phase had not been a part of the study design. Distribution of dementia cases across primary and secondary care data sets are shown in figure 4.

**DISCUSSION**

Medical records allow virtually complete follow-up for dementia research, however there are limitations in terms of accuracy, completeness and underestimation of cases across...
Table 3: Association of sociodemographic factors and dementia in the EPIC-Norfolk cohort in the three data sources combined and individually, mutually adjusted for age, sex, education and social class.

<table>
<thead>
<tr>
<th>Factor</th>
<th>All three data sources (dementia=2555)</th>
<th>HES only (dementia=2087)</th>
<th>MHDS only (dementia=715)</th>
<th>Mortality only (dementia=1229)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>HR</td>
<td>95% CI</td>
<td>P value</td>
<td>HR</td>
</tr>
<tr>
<td>Age per 5 years</td>
<td>2.19</td>
<td>(2.13 to 2.26)</td>
<td>&lt;0.001</td>
<td>2.23</td>
</tr>
<tr>
<td>Sex (men)</td>
<td>0.98</td>
<td>(0.91 to 1.06)</td>
<td>0.6</td>
<td>1.01</td>
</tr>
<tr>
<td>Education (any qualifications)</td>
<td>0.87</td>
<td>(0.80 to 0.95)</td>
<td>0.002</td>
<td>0.87</td>
</tr>
<tr>
<td>Social class (non-manual)</td>
<td>0.99</td>
<td>(0.91 to 1.08)</td>
<td>0.8</td>
<td>0.98</td>
</tr>
</tbody>
</table>

EPIC, European Prospective Investigation of Cancer; HES, Hospital Episode Statistics; MHDS, mental health data set.
in the MHDS, or reflecting that there was no difference in terms of education as to who accesses mental health services. For the sociodemographic factors examined here, we had few missing data, however for more complex factors where there is likely to be more missing data, the potential for these differences to be greater. Such factors and unrecorded dementia cases in a data source could yield different study results. Studies have highlighted regional variation in rates of diagnosis and reliability of existing data.3 We have demonstrated that this heterogeneity exists even in a single geographical region.

We were also able to draw on the relatively new MHDS from NHS Digital, as yet not widely used in research or described in detail. Where there was concordance in cases for the MHDS and HES data, we found that for most cases, the date of diagnosis was earlier in MHDS by just a few months. This is as expected, as the MHDS will be from referrals made from GP, where diagnosis of dementia is usually initiated by family or individuals themselves.13 All activity relating to patients receiving care for a suspected or diagnosed mental health, learning disability or neurodevelopmental conditions is within scope of this data set. This is a further source of clinical and operational data in the NHS that can be used for purposes other than direct patient care. Although the MHDS covers mental healthcare more comprehensively than the other sources, it is limited as most of the clinical information is not coded, but recorded in text, thus, not lending itself so easily large-scale analyses.7

We found MHDS to be complex, with information on individuals (from referral to final discharge) on contact with secondary mental health services. Each subsequent annual release of the mental health data was wider in scope than the previous version. The data in the final year of follow-up gave the greatest number of dementia cases by ICD code. The sharp increase in diagnosis codes shows how important understanding the policy and practice context is to assess the influences and potential biases inherent in use of routine data sources. Additional reasons include changes in coding practices in mental health services, or the way these data are extracted to include diagnoses as well as service codes. The MHDS has far greater potential to provide more complete estimates of diagnosed dementia in the population in the UK. This data set will need further reviewing for future work.

It is currently not possible to examine primary care records via databases such as the Clinical Practice Research Datalink (CPRD) for this cohort. As we had to rely on GP practices to extract the data, we were restricted, for practical reasons, to link to a subset only. The protocol was time and labour intensive. We did not observe the same level of concordance with hospital records as shown in other larger cohort studies.10 22 Some of this discrepancy was due to patients having died as GP records for people who die are moved swiftly to NHS archives. Due to financial incentivisation, there has been an increase in dementia diagnosis in primary care, although concerns of overdiagnosis have been raised and should be monitored carefully.5 Diagnosis of dementia is usually initiated in a primary care setting and could be considered to be the most complete single source of dementia case ascertainment, but underdiagnosis still exists.15 25 29 for various reasons including reluctance of GPs to diagnose dementia.20

As with the UK Biobank study,22 we also found significant proportion of dementia cases in primary care records, that had not been found in hospital or mortality records. In EPIC-Norfolk, this figure was 42% (87 out of the 209 dementia cases identified) compared with 52% reported in UK Biobank. The lower proportion in our study reflects the older age of our cohort, and so more likely to appear in hospital and mortality records than the participants from UK Biobank. UK Biobank included participants from Scotland where data quality is high and centralisation of data, linkable health service data sets make data more accessible to research. This position is not currently shared by the other countries of UK, although may change in the near future with the establishment of Health Data Research UK (HDR-UK) the vision to improve population health, address health inequalities and to drive efficient service provision.30

Figure 4  Distribution of cases (n=209) identified from primary and secondary care in subpopulation of EPIC-Norfolk participants (n=4668). EPIC, European Prospective Investigation of Cancer; GP, general practitioners; HES, Hospital Episode Statistics; MCI, Mild Cognitive Impairment; MH, mental healthcare; NHS, National Health Service.
Another large cohort study, the Million Women Study, found NHS hospital admission data to agree with primary care records. This study only consisted of women, and also like UK Biobank, participants were younger.

The study in the subpopulation of the EPIC-Norfolk cohort with GP records has highlighted several key issues. Even though previous work has shown that practices are confident in identifying dementia cases, the complexity of ‘missingness’ is well recognised. If we had relied on the initial data extract from practices, we would have missed almost half the dementia cases. These cases were only ascertained via the confirmation step which involved a more detailed interrogation of the medical records by the research team. This may be due to lack of technical capacity or capability and could be a further reason for underascertainment. This reveals the potential scale of missed outcomes where known cases are not reflected in extraction of routine records, influencing studies using databases such as CPRD. These findings highlight the importance of extending further training and knowledge of dementia coding in primary care.

It is clear from the characteristics and lower dementia incidence that the subpopulation was different from the overall cohort with regard to age, education and social class and therefore assuming generalisability to wider populations requires appropriate caveats. Although, there were no specific criteria set in approaching practices, the practices that responded were mainly city practices, and the patient-base in these practices was younger, more educated and of higher social class. It is likely that the more research-active practices responded. This bias is clearly reflected in the dementia rates, which was 10% in the EPIC-Norfolk cohort overall, compared with 4% in the subset from the nine GP practices that finally took part.

Whether secondary care data add any further to what can be found in GP record could not be explored in our small study. There were 10 additional participants with a record of dementia in cases from secondary care records, although we were unable to confirm these diagnoses as we could not access their records to check on the detailed information. One additional participant had no indication of dementia or dementia-related condition in the GP records despite a HES record of dementia. This may be a more widespread issue and should be explored in future work involving a larger and wider range of practices.

Databases such as CPRD are extremely powerful data sources, but rely on how dementia is coded in primary care. Furthermore, they are not representative of all practices in the UK based on geography and size. Our findings reveal the potential scale of biases influencing studies using larger primary care databases. They also highlight the importance of extending further training and knowledge of dementia coding in primary care. HES and mortality also rely on data provided by practices. Our findings show the extent to which practice selection impact on dementia case ascertainment and thus on measured rates. Therefore, these data should be used and interpreted with caution.

Recently, there have been changes to coding within primary care with the introduction of new codes SNOMED CT. This new coding system will replace Read codes and eventually also be used in secondary care providing clarity and consistency. However, the implementation of this will take several years and the impact of this is unknown. There is also potential in the mental health data that includes other service-related codes that could be utilised for further insight into the level of cognitive impairment in the community. There are several administrative codes that relate to low, moderate and severe cognitive impairment. This information could be used to supplement the diagnostic information and could be useful in ascertaining milder forms of impairment and dementia.

This study provides confidence that identification of cases via record linkage with hospital admission, mortality and primary care data is sufficient for epidemiologic analyses of risk factors of dementia. The reliability of these data for incidence and prevalence rates is more challenging due to variability in ascertainment and diagnostic criteria which may differ over time and in different populations. In summary, it is important to note that different data sources provide different information. Using a single data source would clearly underestimate dementia outcomes, and so drawing from multiple sources is the best approach to maximise dementia ascertainment from routinely collected health records and to enhance generalisability. Researchers must be fully aware of the strengths and limitations of the data sources they use, identifying the potential sources of bias and be transparent in reporting on how these reflect on the accuracy of their findings.
variables in the analysis file as primary variables or confounding factors means that provision of the data set to other researchers without a Data Transfer Agreement would constitute a risk. The contact for data request is: E-mail: EPIC-Norfolk@mrcepid.cam.ac.uk.

Supplemental material
This content has been supplied by the author(s). It has not been vetted by BMJ Publishing Group Limited (BMJ) and may not have been peer-reviewed. Any opinions or recommendations discussed are solely those of the author(s) and are not endorsed by BMJ. BMJ disclaims all liability and responsibility arising from any reliance placed on the content. Where the content includes any translated material, BMJ does not warrant the accuracy and reliability of the translations (including but not limited to local regulations, clinical guidelines, terminology, drug names and drug dosages), and is not responsible for any error and/or omissions arising from translation and adaptation or otherwise.

Open access
This is an open access article distributed in accordance with the Creative Commons Attribution 4.0 Unported (CC BY 4.0) license, which permits others to copy, redistribute, remix, transform and build upon this work for any purpose, provided the original work is properly cited, a link to the licence is given, and indication of whether changes were made. See: https://creativecommons.org/licenses/by/4.0/.

ORCID iD
Shabina Hayat http://orcid.org/0000-0001-9068-8723

REFERENCES
29. Ford E, Sheppard J, Oliver S, et al. Automated detection of patients with dementia whose symptoms have been identified in primary care but have no formal diagnosis: a retrospective case-control study using electronic primary care records. BMJ Open 2021;11:e039248.