Additional file 4

Sensitivity analyses of the diagnosis definitions using linked data

For this analysis we considered the sensitivity and specificity of both the code lists (possible Alzheimer's disease, probable Alzheimer's disease, non-specific dementia, other dementia, and vascular dementia) and the diagnoses (see Table 2 in manuscript) in the CPRD dataset. Linked data from the Office of National Statistics (ONS) death registry and the Hospital Episode Statistics (HES) inpatient dataset was used as the comparators. HES outpatient data was excluded from the sensitivity analysis as it is known that less than 5.0% of patients have diagnosis recorded in this dataset. (1) This analysis is restricted to patients from practices in England with linked data, which is available for 29,362 patients out of the 40,202 included in the study (73.0%).

We defined the four terms, necessary for the calculation of sensitivity and specificity, as follows:

- True positive: the patient has the code in the CPRD and in the linked data
- False positive: the patient does not have the code in the CPRD but it is in the linked data
- True negative: the patient does not have the code in the CPRD or the linked data
- False negative: the patient has the code in the CPRD but not in the linked data

The sensitivity and specificity could then be calculated as follows (2):

$$\text{Sensitivity} = \frac{\text{true positives}}{\text{true positives} + \text{false negatives}}$$

$$\text{Specificity} = \frac{\text{true negatives}}{\text{true negatives} + \text{false positives}}$$

Diagnosis definitions in the CPRD are determined by Read codes, whereas both the ONS death registry and the HES inpatient dataset use codes from the International Statistical Classification of Diseases and Related Health Problems (ICD). The ICD is maintained by the World Health Organization (WHO) and is currently in its 10th revision: ICD-10. To use this data, we created ICD-10 code lists that correspond to the Read code lists used for the CPRD data extract. These ICD-10 code lists can be found in Additional file 5. ICD-10 and Read codes do not map to each other exactly, with ICD-10 codes generally covering multiple Read codes. As we had been conservative and specific with our approach to the Read codes, we included ICD-10 codes on multiple code lists where appropriate. This helped to ensure the scope of our Read code lists was covered when using the less specific ICD-10 codes. For example, the ICD-10 code 'F03' represents 'Unspecified dementia' and includes the following, many of which refer to diagnoses that are not otherwise specified (NOS):

- Presenile dementia NOS
- Presenile psychosis NOS
- Primary degenerative dementia NOS
- Senile dementia NOS
- Senile dementia, depressed or paranoid type
- Senile psychosis NOS

There are Read codes for each of the above bullet points and for 'Unspecified dementia'. In our Read code lists, we assigned the codes 'Unspecified dementia' and 'Primary degenerative dementia NOS' to the non-specific dementia code list and the remaining codes to the possible Alzheimer's disease code list. We therefore chose to include the ICD-10 code 'F03' on both the non-specific dementia and possible Alzheimer's disease ICD-10 code lists in order to account for the multiple Read codes it relates to.

Tables S4.1 and S4.2 present the sensitivity and specificity of the code lists using the HES inpatient dataset and ONS death registry respectively. Table S4.1 shows that sensitivity using the HES inpatient dataset is best for the code lists non-specific dementia (66.9%) and vascular dementia (61.1%), while sensitivity for other dementia (18.8%) is poor. The code lists relating to Alzheimer's disease are somewhere in the middle with a sensitivity of 41.0% for possible disease and 58.4% for probable disease. On the other hand, the specificity in the HES inpatient dataset is best for the code lists other dementia (98.3%) and vascular dementia (84.5%) and all code lists achieve over 40.0%. Sensitivity indicates whether patients in the CPRD have been included on the right code lists based on the information in the HES inpatient dataset. Specificity indicates whether patients in the CPRD have been excluded from the right code lists based on the same information. The higher specificity of the code lists reflects our conservative approach to the Read code lists, which are used in combination in order to determine diagnosis. As a consequence of this, we expected a lower sensitivity and this is in line with what we observed. While sensitivity is low, the large sample size of our study (that includes patients without linked data) means we have ample power, even if some patients are missed.

Table S4.1: The sensitivity and specificity of the code lists used in the CPRD and HES datasets.

	Patients in CPRD dataset	Patients in HES inpatient dataset	Sensitivity	Specificity
Possible AD	11003	11682	41.0	64.8
Probable AD	8624	6011	58.4	78.1
Non-specific dementia	18369	11632	66.9	40.3
Other dementia	836	1973	18.8	98.3
Vascular dementia	6880	5116	61.1	84.5

AD: Alzheimer's disease; CPRD: Clinical Practice Research Datalink; HES: Hospital Episodes Statistics

Table S4.2 presents the same results but uses the ONS death registry in place of the HES inpatient dataset. In this dataset, we see the best sensitivity for vascular dementia (69.5%) and probable Alzheimer's disease (67.1%). This is likely a reflection of the fact that post mortem examination can help to distinguish between Alzheimer's disease and other types of dementia, which can be difficult to differentiate clinically. On the whole, sensitivity is better using the ONS death registry with a range of 46.7 – 69.5% instead of the 18.8 – 66.9% observed when using the HES inpatient dataset. Specificity remains largely unchanged when using the ONS death registry, though values are slightly lower than those reported for the HES inpatient dataset. The highest specificity remains with other dementia (97.5%) and the lowest with non-specific dementia (38.1%). Much like Table S4.1, Table S4.2 therefore shows a high specificity and low sensitivity for our code lists.

Table S4.2: The sensitivity and specificity of the code lists used in the CPRD and ONS datasets.

	Patients in CPRD dataset	Patients in ONS death registry	Sensitivity	Specificity
Possible AD	11003	4846	46.7	64.3
Probable AD	8624	1911	67.1	73.3
Non-specific dementia	18369	4845	65.7	38.1
Other dementia	836	188	52.1	97.5
Vascular dementia	6880	1298	69.5	78.7

AD: Alzheimer's disease; CPRD: Clinical Practice Research Datalink; ONS: Office of National Statistics

Tables S4.3 and S4.4 present the sensitivity and specificity of the diagnoses using the HES inpatient dataset and ONS death registry respectively. Sensitivity of the diagnoses using the HES inpatient dataset is poor (Table S4.3). For some of the diagnoses, this is due to the small number of patients that have linked data – for example, there are only 29 patients in the CPRD recorded with mixed dementia that excludes Alzheimer's disease and only 384 patients in the HES inpatient dataset. Further to this, the diagnoses rely on multiple code lists. We would therefore expect poor sensitivity in the diagnoses that use multiple code lists with poor sensitivity. This is reflected in the fact that the diagnoses vascular dementia (61.2%) and possible Alzheimer’s disease (59.4%), which rely only on their respective code lists, are performing better than other diagnoses. Specificity using the HES inpatient dataset remains high for diagnoses, with four diagnoses achieving more than 95.0% for this measure and the lowest value being 74.5% for probable Alzheimer’s disease. The high specificity and lower sensitivity of the diagnoses observed here remains in line with the code list analysis.

Table S4.3: The sensitivity and specificity of the diagnoses used in the CPRD and HES datasets.

	Patients in CPRD dataset	Patients in HES inpatient dataset	Sensitivity	Specificity
Possible AD	8069	5007	59.4	79.1
Probable AD	8259	6461	37.3	74.5
Vascular dementia	5429	2016	61.2	84.7
Other dementia	634	465	31.2	98.3
Mixed including possible AD	555	1004	8.3	98.3
Mixed including probable AD	1030	2325	8.9	97.0
Mixed excluding AD	29	384	1.3	99.9
Undiagnosed dementia	5357	12	33.3	81.8

AD: Alzheimer’s disease; CPRD: Clinical Practice Research Datalink; HES: Hospital Episodes Statistics

The sensitivity of the diagnoses using the ONS death registry, presented in Table S4.4, varies but is particularly strong for possible Alzheimer’s disease (65.5%) and vascular dementia (61.6%). As observed in the HES inpatient dataset, sensitivity is poor due to small samples for some diagnoses. Undiagnosed dementia is not recorded in the ONS death registry, which is in line with the definition of this diagnosis, as it relies on non-specific dementia codes that should not be recorded post mortem. As for the codes lists, the sensitivity using the ONS death registry outperforms the equivalent using the HES inpatient dataset for diagnoses. The specificity of diagnoses using the ONS death registry follows much the same pattern as that observed for diagnoses using the HES inpatient dataset. Again, we have multiple diagnoses achieving specificity in excess of 95.0% and the lowest value is recorded for probable Alzheimer’s disease at 73.4%. Once again, we have replication of the sensitivity and specificity scores observed for the code lists. From this, we can conclude that the code lists and ultimately the diagnoses have a high specificity and low sensitivity and this is likely a reflection of the conservative approach we initially took with the code lists.

Table S4.4: The sensitivity and specificity of the diagnoses used in the CPRD and ONS datasets.

	Patients in CPRD dataset	Patients in ONS death registry	Sensitivity	Specificity
Possible AD	8069	1863	65.5	75.1
Probable AD	8259	4752	36.0	73.4
Vascular dementia	5429	1186	61.6	83.3
Other dementia	634	148	44.6	98.1
Mixed including possible AD	555	48	16.7	98.1
Mixed including probable AD	1030	45	8.9	96.5
Mixed excluding AD	29	29	6.9	99.9
Undiagnosed dementia	5357	0	NA	81.8

AD: Alzheimer’s disease; CPRD: Clinical Practice Research Datalink; ONS: Office of National Statistics

References

1. Medicine & Healthcare products Regulatory Agency, National Institute for Health Research, Clinical Practice Research Datalink. Hospital Episode Statistics (HES) Outpatient Care and GOLD Documentation (Set 12). 2016.

2. Lalkhen AG, McCluskey A. Clinical tests: sensitivity and specificity. Contin Educ Anaesth Crit Care Pain. 2008 Dec 1;8(6):221–3.