

# BMJ Open

BMJ Open is committed to open peer review. As part of this commitment we make the peer review history of every article we publish publicly available.

When an article is published we post the peer reviewers' comments and the authors' responses online. We also post the versions of the paper that were used during peer review. These are the versions that the peer review comments apply to.

The versions of the paper that follow are the versions that were submitted during the peer review process. They are not the versions of record or the final published versions. They should not be cited or distributed as the published version of this manuscript.

BMJ Open is an open access journal and the full, final, typeset and author-corrected version of record of the manuscript is available on our site with no access controls, subscription charges or pay-per-view fees (<http://bmjopen.bmj.com>).

If you have any questions on BMJ Open's open peer review process please email [info.bmjopen@bmj.com](mailto:info.bmjopen@bmj.com)

# BMJ Open

## Benefits and limitations of using individual and different combinations of linked English routine data sources in cancer epidemiology studies

Journal:	<i>BMJ Open</i>
Manuscript ID	bmjopen-2020-037719
Article Type:	Original research
Date Submitted by the Author:	13-Feb-2020
Complete List of Authors:	Strongman, Helen; London School of Hygiene and Tropical Medicine, Department of Non-communicable Disease Epidemiology Williams, Rachael; Medicines and Healthcare Products Regulatory Agency, Clinical Practice Research Datalink (CPRD) Bhaskaran, Krishnan; London School of Hygiene & Tropical Medicine, Non-Communicable Disease Epidemiology
Keywords:	ONCOLOGY, EPIDEMIOLOGY, STATISTICS & RESEARCH METHODS, Health informatics < BIOTECHNOLOGY & BIOINFORMATICS

SCHOLARONE™  
Manuscripts



I, the Submitting Author has the right to grant and does grant on behalf of all authors of the Work (as defined in the below author licence), an exclusive licence and/or a non-exclusive licence for contributions from authors who are: i) UK Crown employees; ii) where BMJ has agreed a CC-BY licence shall apply, and/or iii) in accordance with the terms applicable for US Federal Government officers or employees acting as part of their official duties; on a worldwide, perpetual, irrevocable, royalty-free basis to BMJ Publishing Group Ltd ("BMJ") its licensees and where the relevant Journal is co-owned by BMJ to the co-owners of the Journal, to publish the Work in this journal and any other BMJ products and to exploit all rights, as set out in our [licence](#).

The Submitting Author accepts and understands that any supply made under these terms is made by BMJ to the Submitting Author unless you are acting as an employee on behalf of your employer or a postgraduate student of an affiliated institution which is paying any applicable article publishing charge ("APC") for Open Access articles. Where the Submitting Author wishes to make the Work available on an Open Access basis (and intends to pay the relevant APC), the terms of reuse of such Open Access shall be governed by a Creative Commons licence – details of these licences and which [Creative Commons](#) licence will apply to this Work are set out in our licence referred to above.

Other than as permitted in any relevant BMJ Author's Self Archiving Policies, I confirm this Work has not been accepted for publication elsewhere, is not being considered for publication elsewhere and does not duplicate material already published. I confirm all authors consent to publication of this Work and authorise the granting of this licence.

1  
2  
3 1 **Benefits and limitations of using individual and different combinations of linked**  
4  
5  
6 2 **English routine data sources in cancer epidemiology studies**  
7  
8

9 3 Authors: Helen Strongman MSc<sup>1</sup>, Rachael Williams PhD<sup>2</sup>, Prof Krishnan Bhaskaran<sup>1</sup>  
10

11  
12 4 <sup>1</sup> Department of Non-Communicable Diseases Epidemiology, London School of Hygiene and Tropical  
13 5 Medicine, London;

14  
15 6 <sup>2</sup> Clinical Practice Research Datalink, Medicines and Healthcare Products Regulatory Agency,  
16 7 London, UK  
17  
18

19 8

20 9 Correspondence: Helen Strongman, Dept of Non-Communicable Disease Epidemiology, London  
21

22  
23 10 School of Hygiene and Tropical Medicine, London WC1E 7HT, [helen.strongman@lshtm.ac.uk](mailto:helen.strongman@lshtm.ac.uk),  
24

25 11 +44(0)20 7636 8636  
26  
27

28 12

29  
30  
31 13 Abstract: 296 words  
32

33  
34 14 Manuscript: 3352 words  
35

36  
37 15 Tables: 1  
38

39  
40 16 Figures: 5  
41

42  
43 17 Keywords: Cancer; Data Quality; Data Sources; Data Linkage; Epidemiologic Research Designs  
44

45 18 **Abstract**  
46

47  
48  
49 19 **Objectives**  
50

51 20 We aimed to describe the benefits and limitations of using individual and different combinations of  
52  
53 21 linked English electronic health data to identify incident cancers.  
54  
55

56  
57 22 **Design and setting**  
58  
59  
60

1  
2  
3 23 Our descriptive study uses linked Clinical Practice Research Datalink primary care; cancer  
4  
5 24 registration; hospitalisation and death registration data.  
6  
7

## 8 25 **Participants and measures**

9  
10  
11 26 We implemented alternative case definitions to identify first site-specific cancers at the 20 most  
12  
13 27 common cancer sites, based on the first ever cancer diagnosis recorded in each individual data  
14  
15 28 source between 2000-2014, and using commonly used combinations of data sources.  
16  
17

18  
19 29 We calculated positive predictive values and sensitivities of each case definition, compared to a gold  
20  
21 30 standard algorithm that used information from all linked datasets to identify first cancers. We  
22  
23 31 described completeness of grade and stage information in the cancer registration dataset.  
24  
25

## 26 32 **Results**

27  
28  
29 33 168634 gold standard cancers were identified. Positive predictive values of all case definitions were  
30  
31 34  $\geq 94\%$  for the four most common cancers (breast, lung, colorectal, prostate) and  $\geq 80\%$  across cancer  
32  
33 35 sites.  
34  
35

36 36 Sensitivity for case definitions that used cancer registration alone or in combination was  $\geq 92\%$  for  
37  
38 37 the four most common cancers and  $\geq 80\%$  across all cancer sites except bladder cancer (sensitivity  
39  
40 38 65% using cancer registration alone). For case definitions using linked primary care, hospitalisation  
41  
42 39 and death registration data in combination, sensitivity was  $\geq 89\%$  for the four most common cancers,  
43  
44 40 and  $\geq 80\%$  for all cancer sites except kidney (69%), oral cavity (76%) and ovarian cancer (78%).  
45  
46

47 41 Sensitivities were generally lower when primary care or hospitalisation data were used alone.  
48  
49

50 42 Completeness of staging data in cancer registration data was high from 2012.  
51

## 52 43 **Conclusions**

53  
54  
55 44 Ascertainment of incident cancers was good when using cancer registration data alone or in  
56  
57 45 combination with other datasets, and when using a combination of primary care, hospitalisation and  
58  
59 46 death registration data, with variation between cancer sites.  
60

## 47 **Article Summary**

### 48 **Strengths and limitations of the study**

- 49 - We developed a gold standard algorithm using all available data from multiple linked  
50 electronic health data sources in England to identify cases of the 20 most common incident  
51 cancers.
- 52 - Using our gold standard algorithm as a comparator, we then estimated both positive  
53 predictive values and sensitivity values for a range of different pragmatic case definitions for  
54 identifying cancers, using single and multiple data sources.
- 55 - We described similarities and differences in values between age groups, sexes and calendar  
56 years, and the impact of choice of source(s) on mortality rates.
- 57 - We additionally described completeness of stage and grade in cancer registration data.
- 58 - Our research used English data collected between 2000 and 2014 and may not be  
59 generalisable to other countries and time periods.

## 60 **Introduction**

61 The Clinical Practice Research Datalink provides de-identified primary care data linked to additional  
62 secondary health data sources, under a well-governed framework<sup>1</sup>. Use of linked data helps  
63 researchers to answer more epidemiological questions and increase study quality through improved  
64 exposure, outcome and covariate classification<sup>2</sup>. In the field of cancer epidemiology, CPRD primary  
65 care data linked to Hospital Episode Statistics Admitted Patient Care data (HES APC), Office of  
66 National Statistics (ONS) mortality, and National Cancer Registration and Analysis Service (NCRAS)  
67 cancer registration data are used to analyse factors contributing to the risk of cancer and the  
68 consequences of cancer and its treatment. Use of linked data reduces sample size and has cost and  
69 logistical implications, which are greatest for NCRAS data. Research teams therefore commonly  
70 choose not to use all available linked data<sup>3</sup>. Cancer epidemiology studies can also be conducted

1  
2  
3 71 using NCRAS and HES APC data provided by NHS Digital and Public Health England (PHE), without  
4  
5 72 linkage to CPRD primary care data<sup>4</sup>. This provides national coverage at the expense of the detailed  
6  
7 73 health data that are available in primary care records.  
8  
9

10 74 Validation studies assessing concordance between CPRD GOLD, HES APC and NCRAS data have  
11  
12 75 estimated high Positive Predictive Values (PPVs) for CPRD GOLD data and varying proportions of  
13  
14 76 registered cancers that are not captured in CPRD GOLD and HES APC<sup>5-7</sup>. These studies have focused  
15  
16 77 on the most common cancers and concordance between CPRD GOLD only and NCRAS, and do not  
17  
18 78 provide a complete assessment of the benefits and limitations of using different combinations of  
19  
20 79 data sources. National data are available describing completeness of cancer registry data in each  
21  
22 80 collection year<sup>8</sup> and over time for all cancers combined<sup>4</sup>; missingness for individual years has been  
23  
24 81 associated with age, comorbidities and Clinical Commissioning Groups<sup>9,10</sup>.  
25  
26  
27  
28

29 82 We aim to describe and compare the benefits and limitations of using different combinations of  
30  
31 83 linked CPRD primary care data, HES APC, ONS mortality, and NCRAS cancer registration data, for  
32  
33 84 conducting cancer epidemiology studies. Our analyses focus on incident cancer ascertainment as it is  
34  
35 85 a common and important outcome in cancer epidemiology, and it is more difficult to distinguish  
36  
37 86 between secondary, recurrent and primary cancers at a second site in these datasets. We have  
38  
39 87 compared definitions of the twenty most common cancers based on the first ever cancer recorded in  
40  
41 88 individual or combinations of datasets with a gold standard definition comparing information from  
42  
43 89 all four datasets. We also describe the availability of stage, grade and treatment variables over time  
44  
45 90 in the cancer registration data for the CPRD linked cohort. This reflects real life study design and will  
46  
47 91 help researchers to decide which combination of data sources to use for future studies.  
48  
49  
50  
51

52 92  
53  
54

## 55 93 **Methods**

### 56 57 58 94 **Study design and setting** 59 60

1  
2  
3 95 We completed a concordance study using linked CPRD GOLD, HES APC, ONS mortality and NCRAS  
4  
5 96 data (January 2017 CPRD build, set 13 linkage data, study period 1 Jan 2000 – 31 December 2014).  
6  
7  
8 97 The CPRD GOLD database includes de-identified records from participating general practices in the  
9  
10 98 UK who use INPS Vision software<sup>1</sup>. General practice staff can record cancer diagnoses using Read  
11  
12 99 codes or in free text comments boxes, though the latter are not collected by CPRD. Diagnoses will  
13  
14 100 typically be entered during/following a consultation or from written information that is returned to  
15  
16 101 the practice from secondary care. CPRD GOLD data are linked to HES APC, ONS mortality and NCRAS  
17  
18 102 through a trusted third party for English practices that have agreed to participate in the linkage  
19  
20 103 programme<sup>11</sup>. HES APC data are collected by NHS Digital to co-ordinate clinical care in England and  
21  
22 104 calculate hospital payments<sup>12</sup>. Admissions for and related to cancer diagnoses are recorded using  
23  
24 105 ICD-10 codes. National cancer registration data are collected by NCRAS which is part of Public Health  
25  
26 106 England (PHE)<sup>4</sup>. Data include ICD-10 codes to identify the cancer site and more detailed information  
27  
28 107 such as stage and grade. ONS mortality data includes dates and causes of deaths registered in  
29  
30 108 England, recorded using ICD-10 codes.  
31  
32  
33  
34  
35

### 36 109 **Participants, exposures and outcomes**

37  
38  
39 110 Our underlying study population included male and female patients registered in CPRD GOLD  
40  
41 111 practices who were eligible for linkage to HES APC, NCRAS and ONS mortality data and had at least  
42  
43 112 366 days of follow-up between 1 January 1999 and 31 December 2014. Start of follow-up was  
44  
45 113 defined as the latest of the current registration date within the practice and the practice up-to-  
46  
47 114 standard date, and end of follow-up as the earliest of the patient transfer out date, CPRD derived  
48  
49 115 death date, or practice last collection date.  
50  
51

52 116  
53  
54 117 *Identification and classification of cancer codes:* We used code lists to classify cancer records in each  
55  
56 118 of CPRD GOLD, HES APC, and ONS mortality data as one of the 20 most common sites, other  
57  
58 119 specified cancers, history of cancer, secondary cancers, benign tumours, administrative cancer  
59  
60



1  
2  
3 120 codes, unspecified and incompletely specified cancer codes

4  
5 121 (<https://doi.org/10.17037/data.00001519>). Incompletely specified cancer codes could be mapped to

6  
7 122 >1 cancer site (e.g. ICD10 code C68.9 “Malignant neoplasms of urinary organ unspecified” was

8  
9 123 considered consistent with both bladder and kidney cancer). For NCRAS, we accessed coded records

10  
11 124 for the 20 most common cancers. We included cancers recorded in the clinical or referral file for

12  
13 125 CPRD GOLD, cancers recorded in any diagnosis field for HES APC, and the underlying or most

14  
15 126 immediate cancer cause of death in ONS mortality data.

16  
17 127 *Cancer case definitions based on individual sources and combinations of sources:* We developed

18  
19 128 alternative cancer case definitions mirroring those commonly used in epidemiology studies, based

20  
21 129 on identifying the first malignant cancer (excluding administrative codes and benign tumours)

22  
23 130 recorded in various combinations of data sources (NCRAS alone; NCRAS and HES APC; all sources;

24  
25 131 CPRD GOLD, HES APC and ONS mortality; CPRD GOLD alone, HES APC alone). Multiple malignant

26  
27 132 cancers recorded on the index date in CPRD GOLD or HES APC were reclassified as multiple-site

28  
29 133 cancer and were not considered as individual-site cancer records for positive predictive value and

30  
31 134 sensitivity calculations; multiple codes recorded in different sources on the same date were

32  
33 135 reclassified as the site identified in the NCRAS data if available and as multiple-site cancer if not. For

34  
35 136 each case definition, we only examined the first malignant cancer per individual where this occurred

36  
37 137 within the study period and at least one year after the start of follow-up.

38  
39 138 *Gold standard cancer case definition:* We developed a gold standard algorithm that classifies

40  
41 139 incident records of the 20 most common cancers by comparing the first malignant cancer identified

42  
43 140 in each individual source (Figure 1). Cancers recorded in NCRAS alone with no contradictions were

44  
45 141 considered true cases whereas cancers recorded in HES APC alone or GOLD alone required internal

46  
47 142 confirmation within that source in the form of another code for cancer consistent with the same site

48  
49 143 (or with site unspecified) within 6 months and no contradictory codes (e.g. for cancers at other sites)

50  
51 144 in this period. Where cancer records were present in >1 data source, we considered a site-specific

52  
53  
54  
55  
56  
57  
58  
59  
60

1  
2  
3 145 cancer to be a true case (a) if it was recorded as the first cancer in NCRAS and the total number of  
4  
5 146 data sources with records for cancer at that site was equal to or greater than the number of data  
6  
7 147 sources with contradictory records (i.e. records for first cancers at different sites); or (b) where the  
8  
9 148 cancer was not present in NCRAS, if there were more data sources in total with records for cancer at  
10  
11  
12 149 that site than data sources with contradictory records.

13  
14  
15 150 We used NCRAS data to identify stage, grade and treatment where available in the cancer registry  
16  
17 151 only cohort. Binary surgery, chemotherapy and radiotherapy variables were derived using individual  
18  
19 152 records of treatment from the first year after diagnosis.

### 22 153 **Statistical analysis**

23  
24  
25 154 For each cancer site and each individual or combined data source, we combined our applied study  
26  
27 155 definitions with our gold standard definition to classify each applied study definition as a true  
28  
29 156 positive, false positive, or false negative record.

30  
31  
32 157 We used these categories to calculate sensitivity and positive predictive value overall and stratified  
33  
34 158 by age categories (<60, 60-79, 80+), calendar year and sex. We calculated differences in diagnosis  
35  
36 159 dates for true positives by subtracting the gold standard index date from the index date for each  
37  
38 160 source and combination of sources.

39  
40  
41  
42 161 We used Kaplan-Meier methods to describe mortality over time for cancers identified using each  
43  
44 162 definition. The CPRD derived death date was used for these analyses.

45  
46  
47 163 We used the NCRAS only definition to calculate proportions of patients with complete stage and  
48  
49 164 grade and recorded cancer treatment modalities over time.

### 52 165 **Patient public involvement**

53  
54  
55 166 Patients and the public were not involved in conceiving, designing or conducting this study and will  
56  
57 167 not be consulted regarding the dissemination of study results.  
58  
59  
60

## 168 Results

169 Of 14 747 047 research quality patients in the CPRD GOLD January 2017 build, 8 893 326 were  
170 eligible for linkage to HES, ONS mortality and NCRAS data in set 13; 6 791 074 of these were male  
171 and female and had at least one year of follow-up between 1 January 1999 and 31 December 2014  
172 and were included in the study population. Using the gold standard algorithm, 166 614 incident  
173 cases of cancer were identified. The number of patients identified with each cancer is presented in  
174 supplementary appendix table 1. Half (50.0%, n=83 217) of these patients were male; 24.3% (40,502)  
175 aged 0-59, 54.0% (89 940) aged 60-79 and 21.7% (36 172) aged 80 or older.

176 Figure 2 presents PPVs for each case definition, comparing the first recorded cancer in each  
177 combination of data sources with the gold standard algorithm. When using NCRAS data alone, 91.0%  
178 to 99.5% of cancers were confirmed by the algorithm; for 19 out of 20 cancer sites, the NCRAS-only  
179 case definition gave the highest PPV. Case definitions using data sources not including NCRAS  
180 generally had lower PPVs, ranging from 79.6% to 97.3% for individual cancer sites. For the four most  
181 common cancers (breast, lung, colorectal, prostate), PPVs were at least 94% for all case definitions.  
182 Minimal differences in PPVs were observed between age groups, years and sexes (supplementary  
183 appendix figures 1 to 3).

184 Figure 3 presents sensitivity values for each case definition. Sensitivity was generally higher for the  
185 case definitions that included NCRAS data (ranging from 81.0 to 98.7% for individual cancer sites  
186 except bladder cancer identified using NCRAS data alone [64.9%], and  $\geq 92\%$  for the four most  
187 common cancers [breast, lung, colorectal, prostate]). Sensitivity was also generally high for  
188 definitions using a combination of CPRD GOLD, HES APC and ONS mortality data (ranging from 69.3  
189 to 96.3%,  $\geq 89\%$  for the four most common cancers). Sensitivity was lower for case definitions that  
190 used CPRD GOLD alone (range 31.3-89.1% for individual cancer sites) or HES APC alone (range 55.8-  
191 92.2%). Sensitivity values for CPRD GOLD and HES APC increased slightly in younger patients and  
192 more recent years; no differences were observed between males and females (supplementary

1  
2  
3 193 appendix figures 4 to 6). Post-hoc analysis suggested that the low sensitivity of CPRD GOLD only  
4  
5 194 definitions for kidney cancer (sensitivity 31.3%, n false negatives 2901) was driven by missing (n = 1  
6  
7 195 169, 40.3%) or incompletely specified urinary organ cancer codes (n = 1 105, 38.1%) in CPRD GOLD  
8  
9 196 rather than contradictory information about the first cancer record (n = 627, 21.6%). These  
10  
11 197 incompletely specified codes are less likely to be used for bladder cancers (n=85) than kidney  
12  
13 198 cancers (n=1 105). Bladder cancers that were not recorded in NCRAS data (n=3 454) were commonly  
14  
15 199 recorded in both HES APC and CPRD GOLD (n=2 227, 64.5%) or in HES APC only with a subsequent  
16  
17 200 unspecified or bladder cancer record in HES APC within 6 months (n=996, 28.8%).  
18  
19  
20  
21  
22 201 Table 1 describes the number of days (median IQR and 5<sup>th</sup>/95<sup>th</sup> percentile) lag between the date of  
23  
24 202 incident cancers from the gold standard definition and the date of cancer arising from each case  
25  
26 203 definition (i.e. the first record within the specific combinations of data sources used). Case  
27  
28 204 definitions using NCRAS alone and combinations of  $\geq 2$  data sources captured cancers close to the  
29  
30 205 gold standard date (median lag  $\leq 7$  days for all cancer sites), whereas median lags were generally  
31  
32 206 longer for the case definitions using CPRD GOLD alone and HES APC alone.  
33  
34  
35  
36 207 Figure 4 describes mortality over time following incident cancer diagnoses ascertained from each  
37  
38 208 case definition. Minimal differences in mortality were observed between cancers identified from  
39  
40 209 different case definitions. Where variability was observed, cancers identified using CPRD GOLD only  
41  
42 210 had the lowest mortality rates (e.g. kidney cancer) and cancers identified using HES APC only or  
43  
44 211 NCRAS only had higher mortality rates (e.g. prostate cancer and bladder cancer respectively).  
45  
46  
47  
48 212 Figure 5 describes completeness of grade and stage for cancers identified using NCRAS only.  
49  
50 213 Recording of grade was highly variable between cancers with gradual increases in completeness over  
51  
52 214 time. Completeness of staging information was low in earlier calendar years but improved  
53  
54 215 substantially from around 2012 especially for the four most common cancers (min 80.0% 2012,  
55  
56 216 88.6% 2014). Post-hoc logistic regression models adjusted for year and cancer site indicated that  
57  
58 217 completeness of stage and grade were associated with each other and these variables were least  
59  
60

1  
2  
3 218 complete in patients aged  $\geq 80$ ; stage data was more complete for higher grade tumours whereas  
4  
5 219 grade data was more complete for lower stage tumours (supplementary appendix figure 7).  
6  
7  
8 220 Supplementary appendix figure 8 describes recording of treatment modalities identified using  
9  
10 221 NCRAS only. Missing records may indicate that the patient did not receive that treatment modality  
11  
12  
13 222 or that the treatment modality was not recorded.  
14

## 15 223 **Discussion**

### 16 224 *Statement of principal findings*

17  
18  
19 225 We investigated the use of different sources of electronic health record data to identify incident  
20  
21  
22 226 cancers. For all case definitions, using different individual or combined data sources, a minimum of  
23  
24  
25 227 80% of incident site-specific cancers were confirmed using the gold standard algorithm; this rose to  
26  
27  
28 228 94% of the four most common cancers. Use of cancer registration data alone or in any combination  
29  
30  
31 229 of data sources captured at least 80% of site-specific cancers identified by the gold standard  
32  
33 230 algorithm, excepting bladder cancer, and 92.3% of cases for the four most common cancers.  
34  
35 231 Combining all datasets except NCRAS data captured at least 80% of site-specific cancers excepting  
36  
37 232 kidney, oral cavity and ovarian cancers, and captured  $\geq 89\%$  of cases for the four most common  
38  
39 233 cancers. Sensitivity was much more variable when using primary care or hospital data alone, and  
40  
41  
42 234 dropped to 64.9% when identifying bladder cancers using cancer registration data alone. Use of  
43  
44 235 primary care or hospital data alone resulted in a small lag in identifying cancers of interest,  
45  
46 236 compared to the gold standard dates but other case definitions captured cancers close to the gold  
47  
48 237 standard date. Finally, we found that completeness of NCRAS cancer registration stage and grade  
49  
50 238 data increased markedly from 2012 onwards and for specific cancer types; completeness of cancer  
51  
52  
53 239 treatment recording was difficult to assess due to the absence of a missing category.  
54  
55

56 240

### 57 58 59 241 *Strengths and weaknesses of the study*

60

1  
2  
3 242 The main strength of this study is that we have developed a gold standard algorithm using the  
4  
5 243 entirety of the evidence available from CPRD to demonstrate the impact of choice of datasets in  
6  
7 244 identifying incident cancers for real life studies. We have also assessed the value of using NCRAS  
8  
9 245 cancer registration data to measure stage, grade and cancer treatment modalities.  
10  
11  
12 246 A limitation of the study is that our analyses are limited to cancers diagnosed in England between  
13  
14 247 2000 and 2014. We observed minimal changes in PPVs and sensitivities over this time period  
15  
16 248 suggesting that our findings are generalisable to later years. However, substantial improvements in  
17  
18 249 completeness of stage and grade data in 2012 demonstrate that initiatives to improve data can have  
19  
20 250 a profound impact on the quality of data. Another limitation is that our gold standard algorithm pre-  
21  
22 251 weighted NCRAS data as more reliable than other data sources. We feel this is justified as NCRAS is a  
23  
24 252 highly validated data set that matches and merges data from multiple sources<sup>4</sup>. However, this  
25  
26 253 decision will have given case definitions involving NCRAS an inherent advantage in measures of  
27  
28 254 positive predictive value and sensitivity. The algorithm will also have been affected by different  
29  
30 255 lengths of follow-up data available in the different data sources. For example, NCRAS data collection  
31  
32 256 started later than CPRD GOLD and HES which may account for some of the misclassification of  
33  
34 257 incident cases when using NCRAS alone. Requiring internal confirmation within 6 months for cancers  
35  
36 258 recorded in HES APC or CPRD GOLD alone in our GOLD standard definition is more likely to discount  
37  
38 259 cancers with poorer prognoses and those recorded in the last 6 months of follow-up. Our data cut  
39  
40 260 only included NCRAS data for the top 20 cancers; earlier cancers at other sites will have been missed  
41  
42 261 in this study.  
43  
44  
45  
46  
47  
48  
49 262 It is also important to note that as the gold standard algorithm uses data recorded after the first  
50  
51 263 record of the cancer site in any source (index date), it cannot be used to identify outcomes in applied  
52  
53 264 studies and follow-up of cohort studies with cancer as an exposure would need to start at least 6  
54  
55 265 months after diagnosis; our first ever cancer record in any source definition would be more  
56  
57 266 appropriate for most studies.  
58  
59  
60

267

268 *Strengths and weaknesses in relation to other studies, discussing important differences in results*

269 The most up to date study describing concordance between linked English datasets demonstrated  
270 that 2-4% of the 5 most common cancers recorded in CPRD are not confirmed in either HES APC or  
271 cancer registration data and 9-33% of registered cancers are not recorded in CPRD GOLD<sup>13</sup>. For  
272 cancers recorded in both sources, the diagnosis date was a median of 6-16 days later in CPRD GOLD  
273 than in the registration data. Using CPRD GOLD alone to identify these cancers marginally over  
274 represented younger, healthier patients and identified 1-6% fewer deaths in the first five years after  
275 diagnosis. Use of HES APC only identified a higher proportion of patients with the correct diagnosis  
276 date than CPRD GOLD, but over represented older patients and those diagnosed through the  
277 emergency route. The majority of registered cancers were picked up using both CPRD GOLD and HES  
278 APC (ranging from 91% for lung cancer to 97% for breast cancer). Previous research demonstrated  
279 similar results with substantial differences between cancer types<sup>5,6</sup>.

280 Our study is consistent with these results and provides more complete evidence for a wide range of  
281 cancers which will allow researchers to understand the strengths and limitations of different study  
282 designs.

283 We have also demonstrated the added value of using cancer registration data to measure stage and  
284 grade of incident cancers from about 2012 onwards. Levels of data completeness of staging  
285 information in the CPRD extract in 2012 were similar to those reported by the United Kingdom and  
286 Ireland Association of Cancer Registries (UKAICR)<sup>8</sup>.

287

288

289 *Meaning of the study: possible explanations and implications for clinicians and policymakers*

1  
2  
3 290 Use of NCRAS cancer registration data maximised the proportion of cases confirmed as true positive  
4  
5 291 based on all available linked information and captured the highest proportion of true positive cases;  
6  
7 292 highly complete staging and grading information is available from this source from approximately  
8  
9 293 2012. Case definitions based on a combination of CPRD GOLD, HES APC and ONS mortality data also  
10  
11 294 had acceptable validity for the majority of cancer sites including the four most common cancers.  
12  
13  
14  
15 295 These findings should be considered when deciding which data sources to include in research studies  
16  
17 296 and which sources to use to define cancer exposures, outcomes and covariates.  
18  
19  
20  
21 297

### 22 23 298 *Unanswered questions and future research*

24  
25  
26 299 Further research is required to understand differences in cancer data recording with CPRD GOLD and  
27  
28 300 CPRD Aurum, CPRD's recently launched primary care database based on records from EMIS  
29  
30 301 practices<sup>14</sup>. Use of NCRAS's recently launched Systemic Anti-Cancer Therapy (SACT)<sup>15</sup> and National  
31  
32 302 Radiotherapy Datasets will also improve ascertainment of therapies for future studies.  
33  
34  
35

### 36 303 *Conclusion*

37  
38  
39 304 Completeness and accuracy of recording of cancers in English data sources is high particularly when  
40  
41 305 using NCRAS cancer registration data alone or in any combination with other data sources, and when  
42  
43 306 using a combination of CPRD GOLD, HES APC and ONS mortality data, with variation between cancer  
44  
45 307 types. Completeness of cancer stage and grade variables in NCRAS was low before 2012 but appears  
46  
47 308 to have substantially improved for most cancers in more recent calendar periods. This study  
48  
49 309 describes likely levels of misclassification for a range of data sources, combinations and cancer sites  
50  
51 310 enabling cancer epidemiologists to optimise study design and better understand the limitations of  
52  
53 311 their research.  
54  
55

### 56 57 312 **Funding**



1  
2  
3 313 CPRD funded access to the linked data sources used in this work. This work was additionally  
4  
5 314 supported by the Wellcome Trust and Royal Society grant number 107731/Z/15/Z.  
6  
7

### 8 315 **Acknowledgements**

9  
10  
11 316 This study is based in part on data from the Clinical Practice Research Datalink obtained under  
12  
13 317 licence from the UK Medicines and Healthcare products Regulatory Agency. The data is provided by  
14  
15 318 patients and collected by the NHS as part of their care and support. The interpretation and  
16  
17 319 conclusions contained in this study are those of the author/s alone.  
18  
19

### 20 21 320 **Protocol**

22  
23  
24 321 Available on request  
25  
26

### 27 322 **Competing Interests**

28  
29  
30 323 RW is employed by CPRD. HS and KB have academic honorary contracts at PHE for a separate  
31  
32 324 collaborative research study.  
33  
34

### 35 325 **Contributions**

36  
37  
38  
39 326 All authors conceived the study and contributed to the study design. HS and KB did the data  
40  
41 327 management. HS did the statistical analysis and wrote the first draft. All authors contributed to  
42  
43 328 subsequent drafts.  
44  
45

### 46 329 **Patient consent for publication**

47  
48  
49 330 Not required  
50  
51

### 52 331 **Data sharing**

53  
54  
55  
56 332 Data were obtained from the Clinical Practice Research Datalink, provided by the UK Medicines and  
57  
58 333 Healthcare products Regulatory Agency. The authors' licence for using these data does not allow  
59  
60

1  
2  
3 334 sharing of raw data with third parties. Information about access to Clinical Practice Research  
4  
5 335 Datalink data is available here: <https://www.cprd.com/research-applications>. Code lists for this  
6  
7 336 study are available at <https://doi.org/10.17037/data.00001519>  
8  
9

10 337 **References**  
11  
12

- 13 338 1 Herrett E, Gallagher AM, Bhaskaran K, *et al*. Data Resource Profile: Clinical Practice Research  
14  
15 339 Datalink (CPRD). *Int J Epidemiol* 2015; **44**: 827–36.  
16  
17  
18 340 2 Padmanabhan S, Carty L, Cameron E, Ghosh RE, Williams R, Strongman H. Approach to record  
19  
20 341 linkage of primary care data from Clinical Practice Research Datalink to other health-related  
21  
22 342 patient data: overview and implications. *Eur J Epidemiol* 2019; **34**: 91–9.  
23  
24  
25  
26 343 3 Badrick E, Renehan I, Renehan AG. Linkage of the UK Clinical Practice Research Datalink with  
27  
28 344 the national cancer registry. *Eur J Epidemiol* 2019; **34**: 101–2.  
29  
30  
31 345 4 Henson KE, Elliss-Brookes L, Coupland VH, *et al*. Data Resource Profile: National Cancer  
32  
33 346 Registration Dataset in England. *Int J Epidemiol* 2019; published online April 23.  
34  
35 347 DOI:10.1093/ije/dyz076.  
36  
37  
38 348 5 Dregan A, Moller H, Murray-Thomas T, Gulliford MC. Validity of cancer diagnosis in a primary  
39  
40 349 care database compared with linked cancer registrations in England. Population-based cohort  
41  
42 350 study. *Cancer Epidemiol* 2012; **36**: 425–9.  
43  
44  
45  
46 351 6 Boggon R, van Staa TP, Chapman M, Gallagher AM, Hammad TA, Richards MA. Cancer  
47  
48 352 recording and mortality in the General Practice Research Database and linked cancer  
49  
50 353 registries. *Pharmacoepidemiol Drug Saf* 2013; **22**: 168–75.  
51  
52  
53 354 7 Rañopa M, Douglas I, van Staa T, *et al*. The identification of incident cancers in UK primary  
54  
55 355 care databases: a systematic review. *Pharmacoepidemiol Drug Saf* 2015; **24**: 11–8.  
56  
57  
58 356 8 UK and Ireland Association of Cancer Registries. UK and Ireland Association of Cancer  
59  
60

- 1  
2  
3 357 Registries. <http://www.ukiacr.org/kpis/> (accessed March 18, 2019).  
4  
5  
6 358 9 Di Girolamo C, Walters S, Benitez Majano S, *et al*. Characteristics of patients with missing  
7  
8 359 information on stage: a population-based study of patients diagnosed with colon, lung or  
9  
10 360 breast cancer in England in 2013. *BMC Cancer* 2018; **18**: 492.  
11  
12  
13 361 10 Barclay ME, Lyratzopoulos G, Greenberg DC, Abel GA. Missing data and chance variation in  
14  
15 362 public reporting of cancer stage at diagnosis: Cross-sectional analysis of population-based  
16  
17 363 data in England. *Cancer Epidemiol* 2018; **52**: 28–42.  
18  
19  
20  
21 364 11 Padmanabhan S, Smith O, Strongman H. Supplementing linked datasets with meaningful  
22  
23 365 meta-data to enable high quality research. *Int J Popul Data Sci* 2017; **1**.  
24  
25 366 DOI:10.23889/ijpds.v1i1.300.  
26  
27  
28 367 12 Herbert A, Wijlaars L, Zylbersztejn A, Cromwell D, Hardelid P. Data Resource Profile: Hospital  
29  
30 368 Episode Statistics Admitted Patient Care (HES APC). *Int J Epidemiol* 2017; **46**: 1093-1093i.  
31  
32  
33 369 13 Arhi CS, Bottle A, Burns EM, *et al*. Comparison of cancer diagnosis recording between the  
34  
35 370 Clinical Practice Research Datalink, Cancer Registry and Hospital Episodes Statistics. *Cancer*  
36  
37 371 *Epidemiol* 2018; **57**: 148–57.  
38  
39  
40  
41 372 14 Wolf A, Dedman D, Campbell J, *et al*. Data resource profile: Clinical Practice Research Datalink  
42  
43 373 (CPRD) Aurum. *Int J Epidemiol* 2019; published online March 11. DOI:10.1093/ije/dyz034.  
44  
45  
46 374 15 Bright CJ, Lawton S, Benson S, *et al*. Data Resource Profile: The Systemic Anti-Cancer Therapy  
47  
48 375 (SACT) Dataset. *Int J Epidemiol* 2019; published online July 24. DOI:10.1093/ije/dyz137.  
49  
50  
51 376  
52  
53  
54 377  
55  
56  
57  
58  
59  
60

**Table 1:** Time in days from main gold standard diagnosis date to first ever record in each combination of sources

Cancer	NCRAS		NCRAS & HES APC		CPRD GOLD, HES APC & ONS MORTALITY		CPRD GOLD		HES APC	
	median (IQR)	5th-95th percentile	median (IQR)	5th-95th percentile	median (IQR)	5th-95th percentile	median (IQR)	5th-95th percentile	median (IQR)	5th-95th percentile
Oral Cavity (C00-06)	0 (0, 0)	0-20	0 (0, 0)	0-13	0 (0, 18)	0-59	12 (0, 26)	0-80	13 (0, 40)	0-93
Oesophageal (C15)	0 (0, 1)	0-30	0 (0, 0)	0-6	0 (0, 0)	0-30	7 (0, 18)	0-59	0 (0, 6)	0-86
Stomach (C16)	0 (0, 2)	0-27	0 (0, 0)	0-0	0 (0, 0)	0-38	10 (1, 22)	0-63	0 (0, 0)	0-64
Colorectal (C18-C20)*	0 (0, 3)	0-41	0 (0, 0)	0-19	0 (0, 0)	0-37	7 (0, 21)	0-70	0 (0, 16)	0-90
Liver (C22)	0 (0, 7)	0-87	0 (0, 0)	0-52	0 (0, 4)	0-72	9 (0, 29)	0-113	0 (0, 33)	0-174
Pancreas (C25)	0 (0, 8)	0-56	0 (0, 0)	0-23	0 (0, 0)	0-53	9 (0, 22)	0-76	0 (0, 8)	0-103
Lung (C34)*	0 (0, 5)	0-42	0 (0, 0)	0-20	0 (0, 4)	0-56	10 (0, 22)	0-85	0 (0, 19)	0-192
Malignant melanoma (C43)	0 (0, 0)	0-23	0 (0, 0)	0-29	0 (0, 21)	0-67	12 (0, 26)	0-74	31 (0, 62)	0-240
Breast (C50)*	0 (0, 0)	0-26	0 (0, 0)	0-27	7 (0, 14)	0-37	7 (0, 14)	0-48	27 (16, 41)	0-364
Cervix (C53)	0 (0, 0)	0-15	0 (0, 0)	0-3	4 (0, 21)	0-74	13 (5, 28)	0-79	17 (0, 48)	0-113
Uterus (C54-55)	0 (0, 0)	0-19	0 (0, 0)	0-4	0 (0, 19)	0-56	14 (7, 27)	0-69	8 (0, 41)	0-89
Ovaries (C56)	0 (0, 3)	0-33	0 (0, 0)	0-20	0 (0, 0)	0-42	10 (0, 24)	0-96	0 (0, 15)	0-98
Prostate (C61)*	0 (0, 0)	0-68	0 (0, 0)	0-77	3 (0, 22)	0-156	15 (3, 29)	0-113	66 (0, 425)	0-2,108
Kidney (C64)	0 (0, 5)	0-66	0 (0, 0)	0-33	0 (0, 0)	0-97	0 (0, 23)	0-117	0 (0, 19)	0-250
Bladder (C67)	1 (0, 15)	0-220	0 (0, 0)	0-31	0 (0, 0)	0-31	8 (0, 30)	0-149	0 (0, 2)	0-97
Brain/CNS (C71-72)	1 (0, 8)	0-63	0 (0, 0)	0-33	0 (0, 0)	0-33	8 (0, 21)	0-68	0 (0, 2)	0-168
Thyroid (C73)	0 (0, 0)	0-28	0 (0, 0)	0-19	0 (0, 26)	0-89	22 (4, 42)	0-127	4 (0, 59)	0-154
Non-Hodgkin lymphoma (C82-85)	0 (0, 3)	0-43	0 (0, 0)	0-32	0 (0, 12)	0-62	16 (4, 32)	0-118	0 (0, 31)	0-547
Multiple myeloma (C90)	0 (0, 8)	0-235	0 (0, 0)	0-80	0 (0, 2)	0-78	11 (0, 28)	0-147	0 (0, 43)	0-726
Leukemia (C91-95)	0 (0, 7)	0-890	0 (0, 1)	0-1,033	0 (0, 0)	0-92	1 (0, 20)	0-138	1 (0, 196)	0-1,811

Footnote: Number of days between main gold standard diagnosis date and applied definitions. Cancer sites are ordered according to corresponding codes from the International Classification of Diseases, version 10

(ICD-10). \*Four most common cancer sites. All sources definition not shown as diagnosis date is the same as the gold standard definition by default. NCRAS = National Cancer Registration and Analysis Service cancer

registration data. CPRD = Clinical Practice Research Datalink. HES APC = Hospital Episode Statistics Admitted Patient Care data. ONS = Office for National Statistics

1  
2  
3 **Figure 1:**  
4  
5

6 Title: Gold standard algorithm to identify incident site-specific cancers using all data sources  
7  
8

9 **Figure 2:**  
10  
11

12 Title: Positive Predictive Value of cancer diagnoses for each combination of sources when compared  
13  
14 to the main gold standard algorithm  
15

16  
17 Legend: Percentage of incident cancers defined using the first ever record in each combination of  
18  
19 sources confirmed by a gold standard algorithm that considers confirmatory and contradictory data  
20  
21 from each source. Cancer sites are ordered according to corresponding codes from the International  
22  
23 Classification of Diseases, version 10 (ICD-10). \*Four most common cancer sites. NCRAS = National  
24  
25 Cancer Registration and Analysis Service cancer registration data. CPRD = Clinical Practice  
26  
27 Research Datalink. HES APC = Hospital Episode Statistics Admitted Patient Care data. ONS = Office  
28  
29 for National Statistics  
30  
31  
32

33 **Figure 3:**  
34  
35

36 Title: Sensitivity of cancer diagnoses for each combination of sources when compared to the main  
37  
38 gold standard algorithm  
39  
40

41 Legend: Percentage of incident cancers identified using the main gold standard algorithm that  
42  
43 considers confirmatory and contradictory data from each source that are identified using the first  
44  
45 ever record in each combination of sources. Cancer sites are ordered according to corresponding  
46  
47 codes from the International Classification of Diseases, version 10 (ICD-10). \*Four most common  
48  
49 cancer sites. NCRAS = National Cancer Registration and Analysis Service cancer registration data.  
50  
51 CPRD = Clinical Practice Research Datalink. HES APC = Hospital Episode Statistics Admitted Patient  
52  
53 Care data. ONS = Office for National Statistics  
54  
55  
56

57  
58  
59 **Figure 4:**  
60

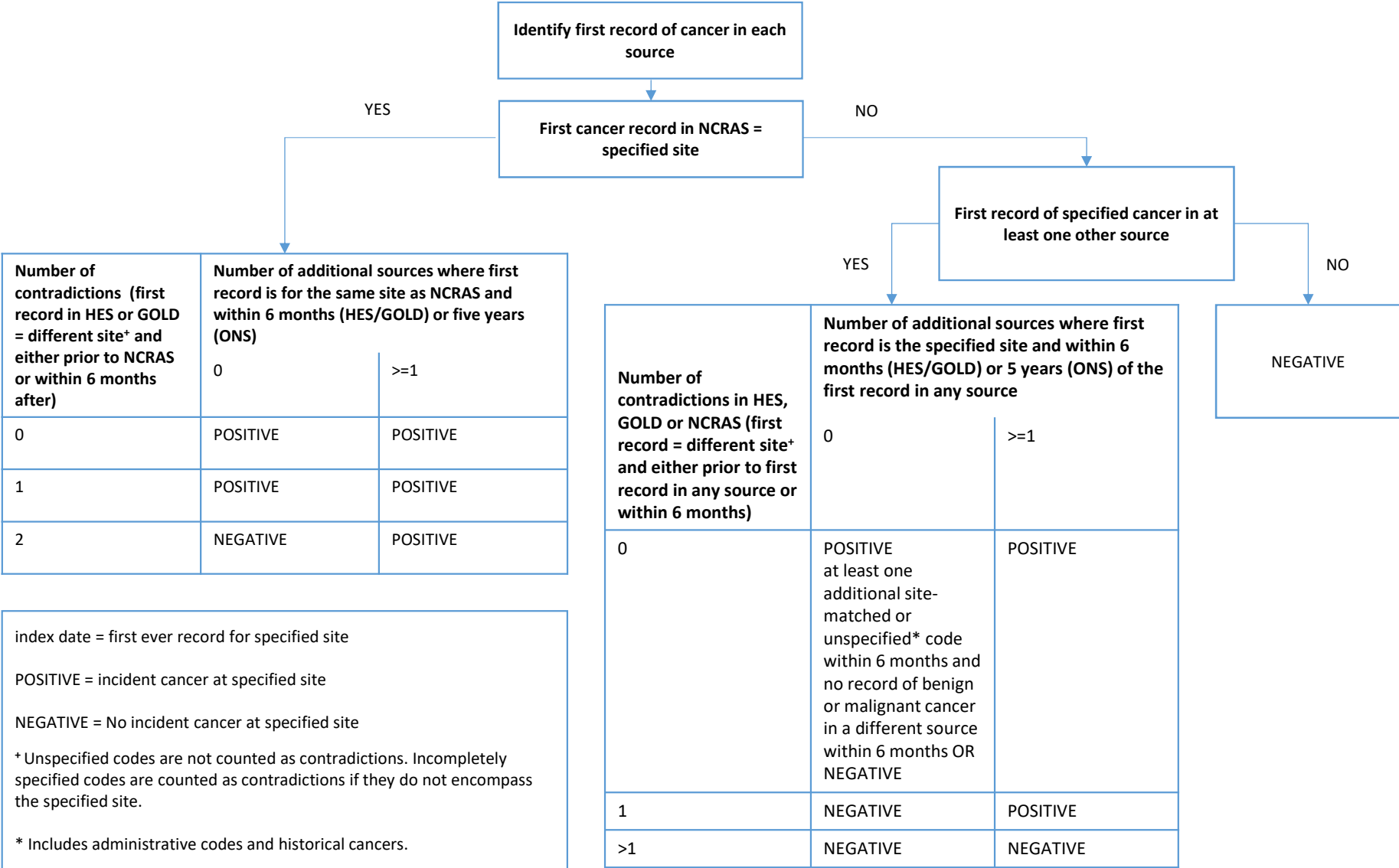
1  
2  
3 Title: Mortality following first ever record of cancer in each combination of sources  
4  
5

6 Legend: Cancer sites are ordered according to corresponding codes from the International  
7  
8 Classification of Diseases, version 10 (ICD-10). \*Four most common cancer sites. NHL = Non hodgkin  
9  
10 lymphoma. NCRAS = National Cancer Registration and Analysis Service cancer registration data.  
11  
12 CPRD = Clinical Practice Research Datalink. HES APC = Hospital Episode Statistics Admitted Patient  
13  
14 Care data. ONS = Office for National Statistics  
15  
16

17  
18 **Figure 5:**

19  
20  
21 Title: Completeness of grade and stage for cancers identified using NCRAS data only  
22

23  
24 Legend: Cancer sites are ordered according to corresponding codes from the International  
25  
26 Classification of Diseases, version 10 (ICD-10). \*Four most common cancer sites. NHL = Non hodgkin  
27  
28 lymphoma.  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60



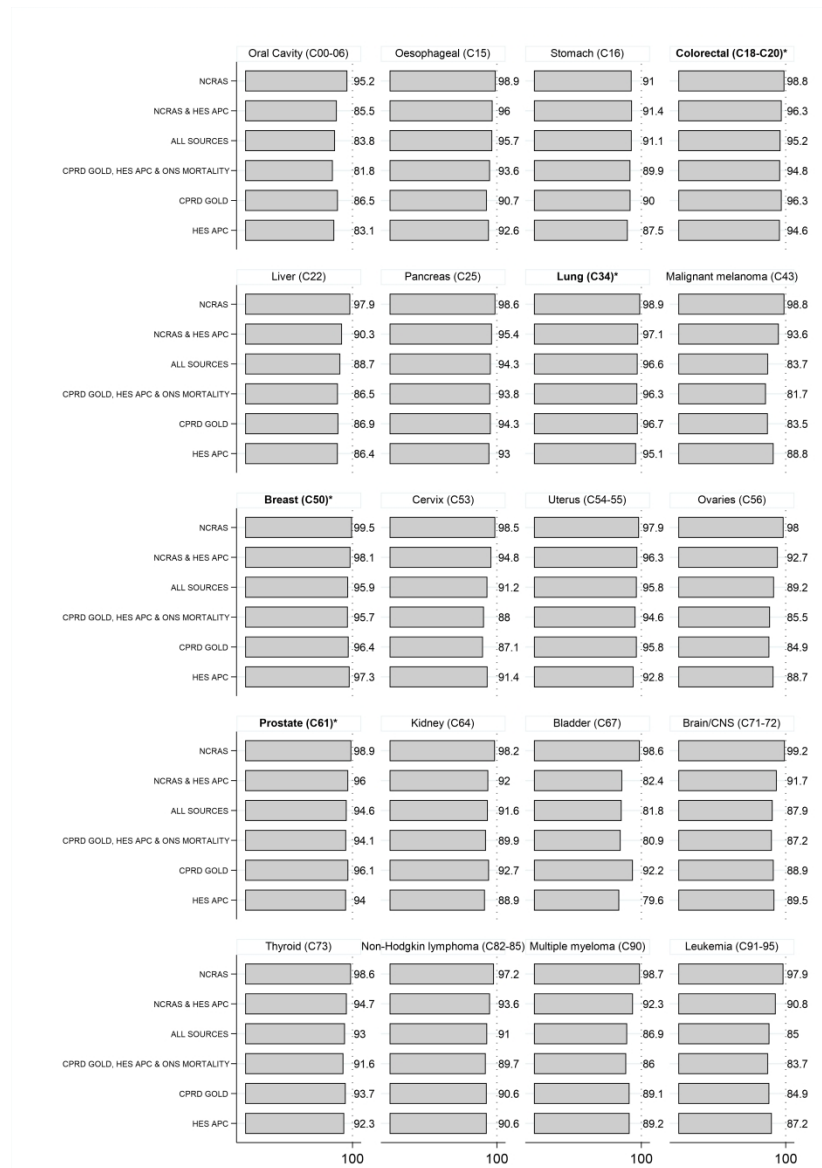
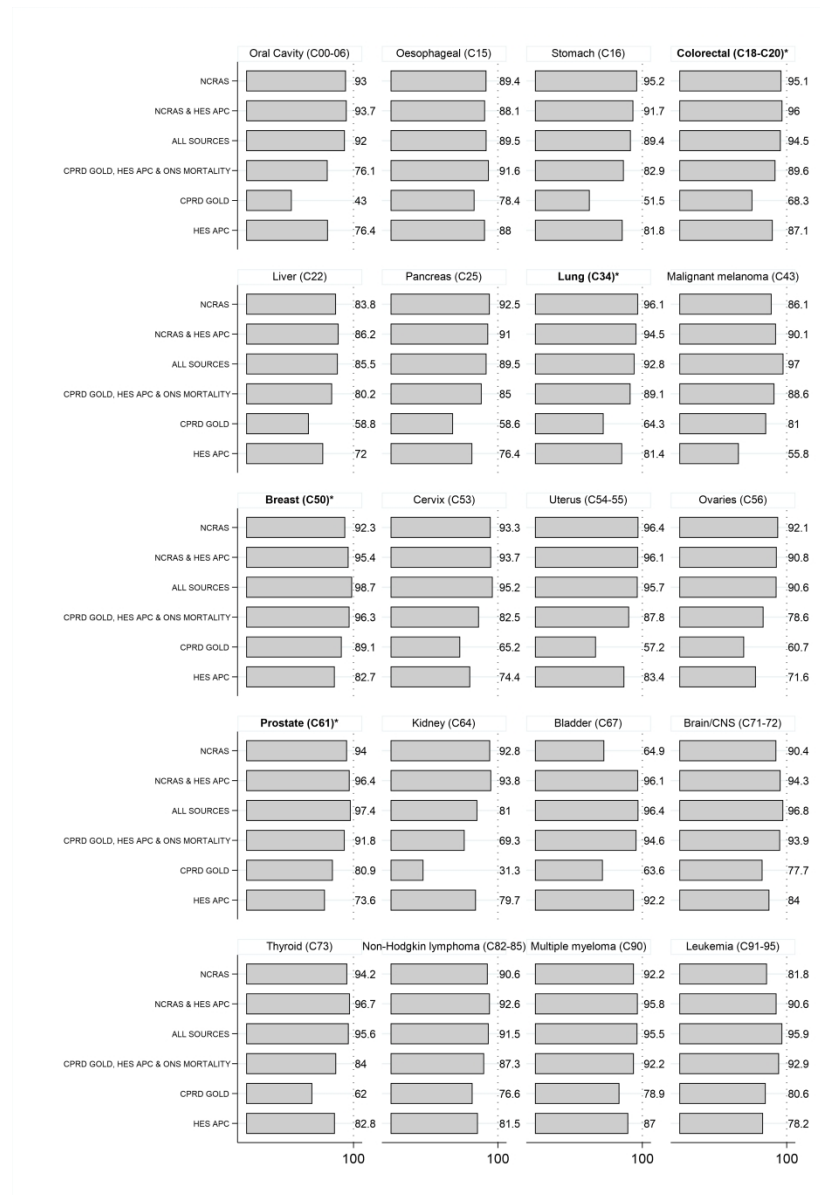


Figure 2: Positive Predictive Value of cancer diagnoses for each combination of sources when compared to the main gold standard algorithm

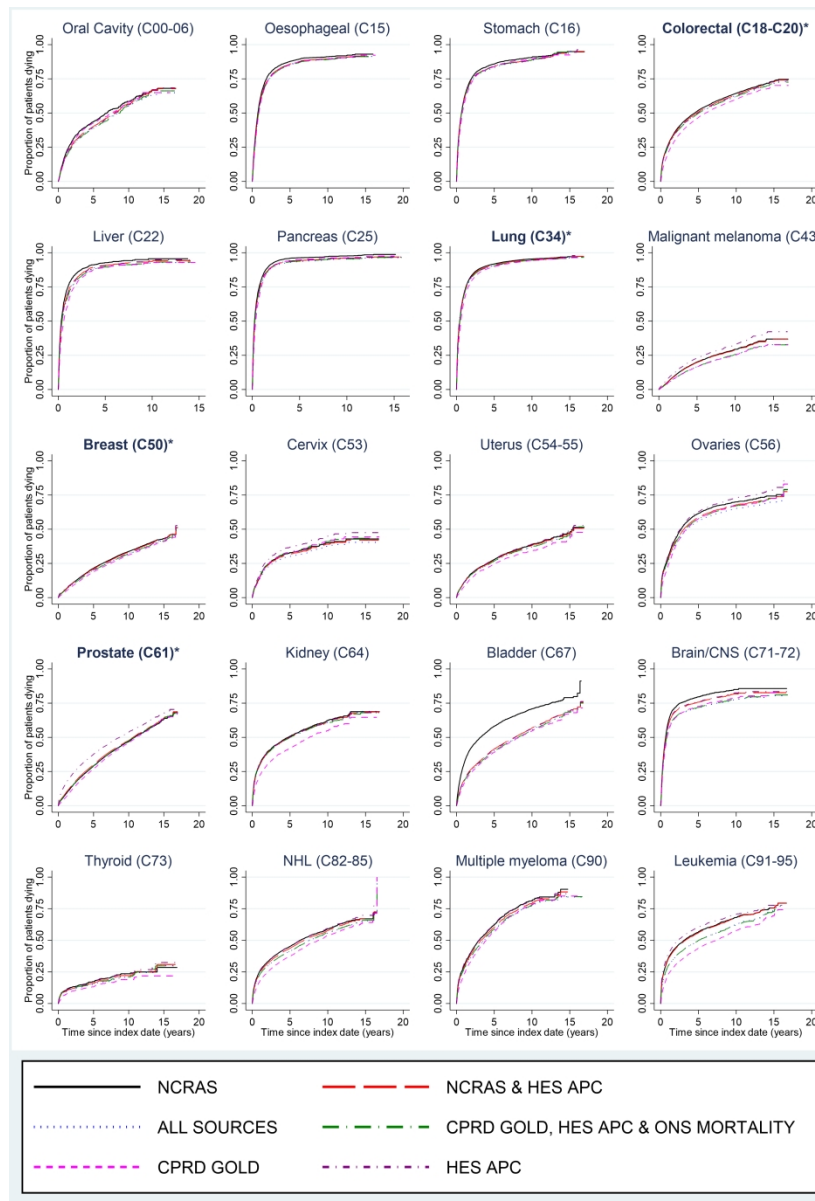
Legend: Percentage of incident cancers defined using the first ever record in each combination of sources confirmed by a gold standard algorithm that considers confirmatory and contradictory data from each source. Cancer sites are ordered according to corresponding codes from the International Classification of Diseases, version 10 (ICD-10). \*Four most common cancer sites. NCRAS = National Cancer Registration and Analysis Service cancer registration data. CPRD = Clinical Practice Research Datalink. HES APC = Hospital Episode Statistics Admitted Patient Care data. ONS = Office for National Statistics



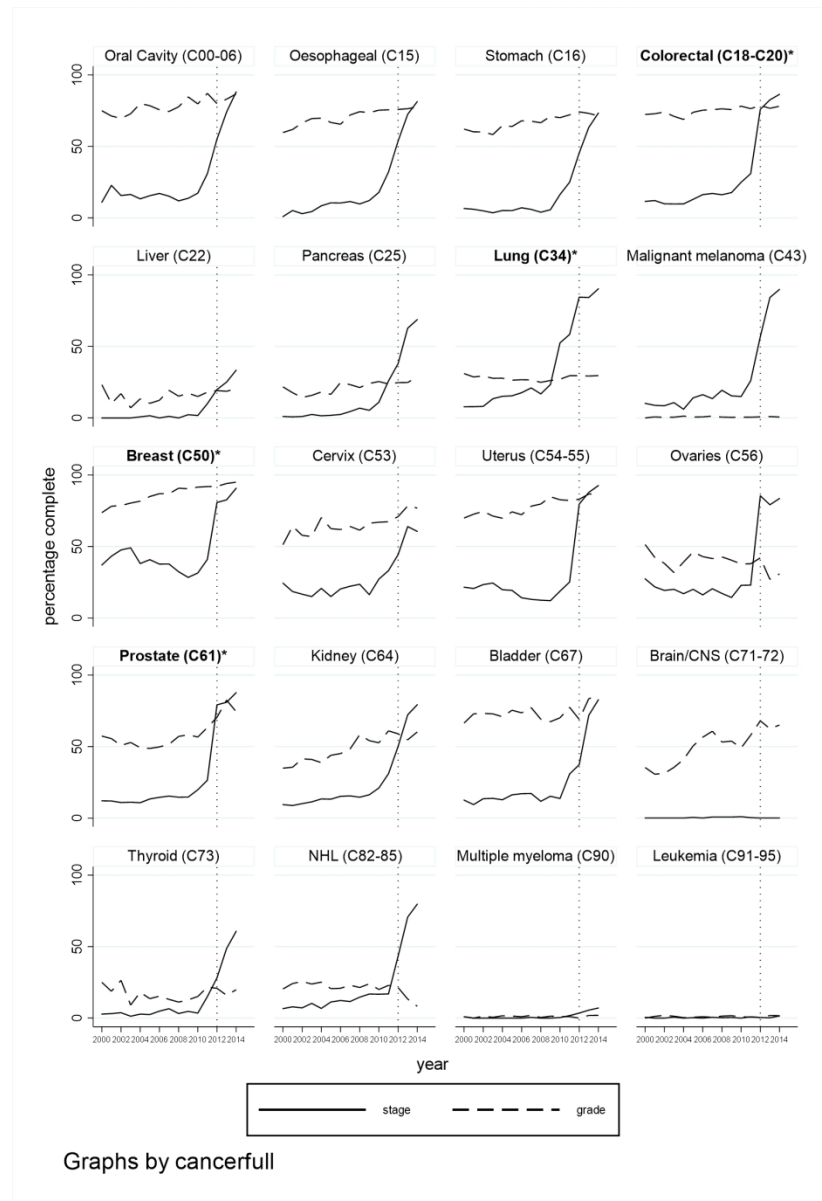


Caption : Figure 3: Sensitivity of cancer diagnoses for each combination of sources when compared to the main gold standard algorithm

Legend: Percentage of incident cancers identified using the main gold standard algorithm that are identified using the first ever record in each combination of sources. Cancer sites are ordered according to corresponding codes from the International Classification of Diseases, version 10 (ICD-10). \*Four most common cancer sites. NCRAS = National Cancer Registration and Analysis Service cancer registration data. CPRD = Clinical Practice Research Datalink. HES APC = Hospital Episode Statistics Admitted Patient Care data. ONS = Office for National Statistics



Caption: Figure 4: Mortality following first ever record of cancer in each combination of sources  
 Legend: Cancer sites are ordered according to corresponding codes from the International Classification of Diseases, version 10 (ICD-10). \*Four most common cancer sites. NHL = Non hodgkin lymphoma. NCRAS = National Cancer Registration and Analysis Service cancer registration data. CPRD = Clinical Practice Research Datalink. HES APC = Hospital Episode Statistics Admitted Patient Care data. ONS = Office for National Statistics



Caption : Figure 5: Completeness of grade and stage for cancers identified using NCRAS data only  
 Legend: Cancer sites are ordered according to corresponding codes from the International Classification of Diseases, version 10 (ICD-10). \*Four most common cancer sites. NHL = Non hodgkin lymphoma.

## Supplementary appendix

### Benefits and limitations of using individual and different combinations of linked English routine data sources in cancer epidemiology studies

Table 1: Number of patients identified with each cancer site using the gold standard algorithm

Cancer site	Number of patients
Oral Cavity (C00-06)	2105
Oesophageal (C15)	5212
Stomach (C16)	4041
Colorectal (C18-C20)*	22276
Liver (C22)	2249
Pancreas (C25)	5048
Lung (C34)	22183
Malignant melanoma (C43)	7286
Breast (C50)	29338
Cervix (C53)	1509
Uterus (C54-55)	4344
Ovaries (C56)	4174
Prostate (C61)	24936
Kidney (C64)	4118
Bladder (C67)	8908
Brain/CNS (C71-72)	2926
Thyroid (C73)	1317
NHL (C82-85)	6669
Multiple myeloma (C90)	2684
Leukemia (C91-95)	5291
Total	166614

Figure 1: Positive Predictive Value by age

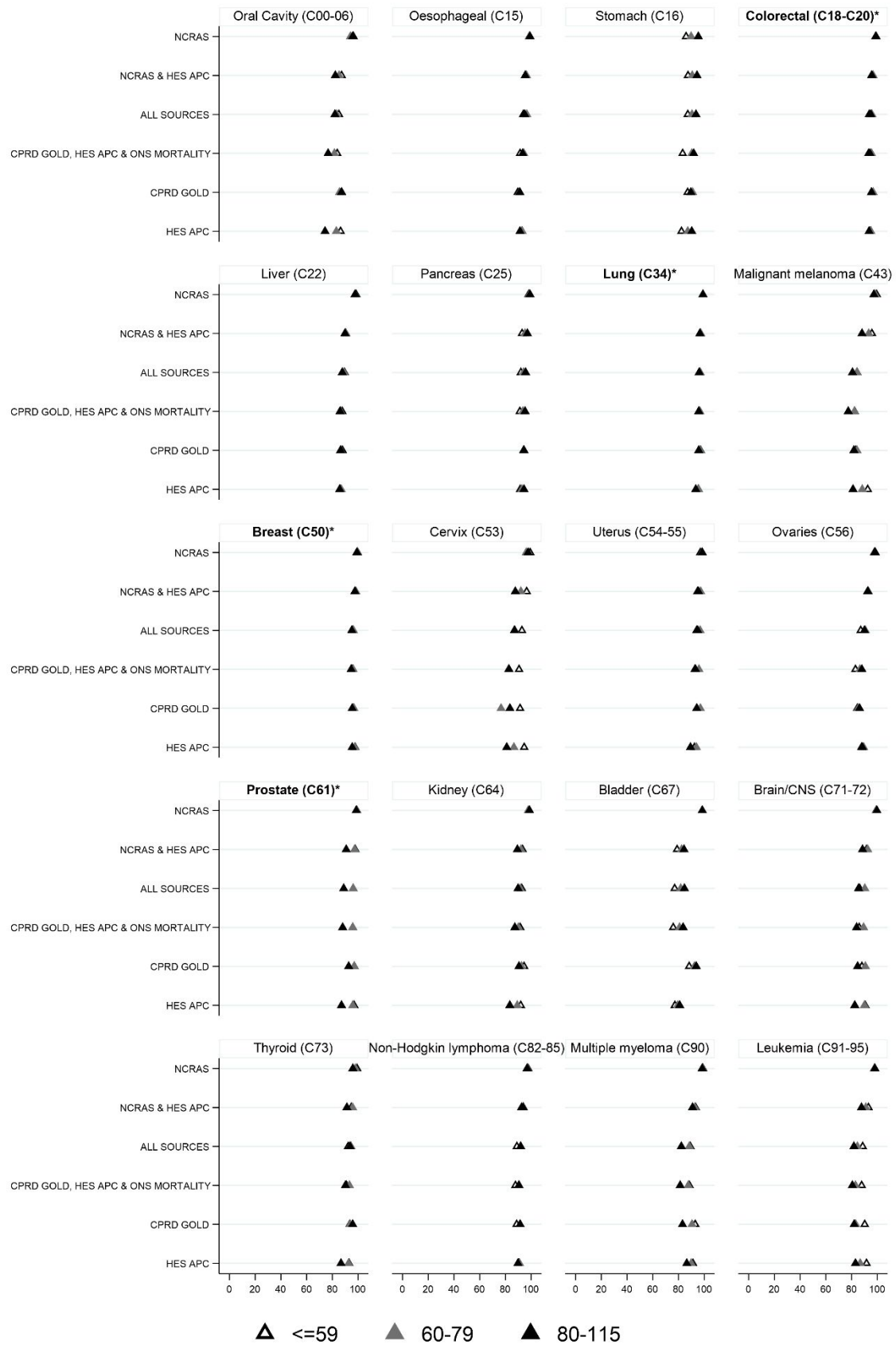


Figure 2: Positive Predictive Value by sex

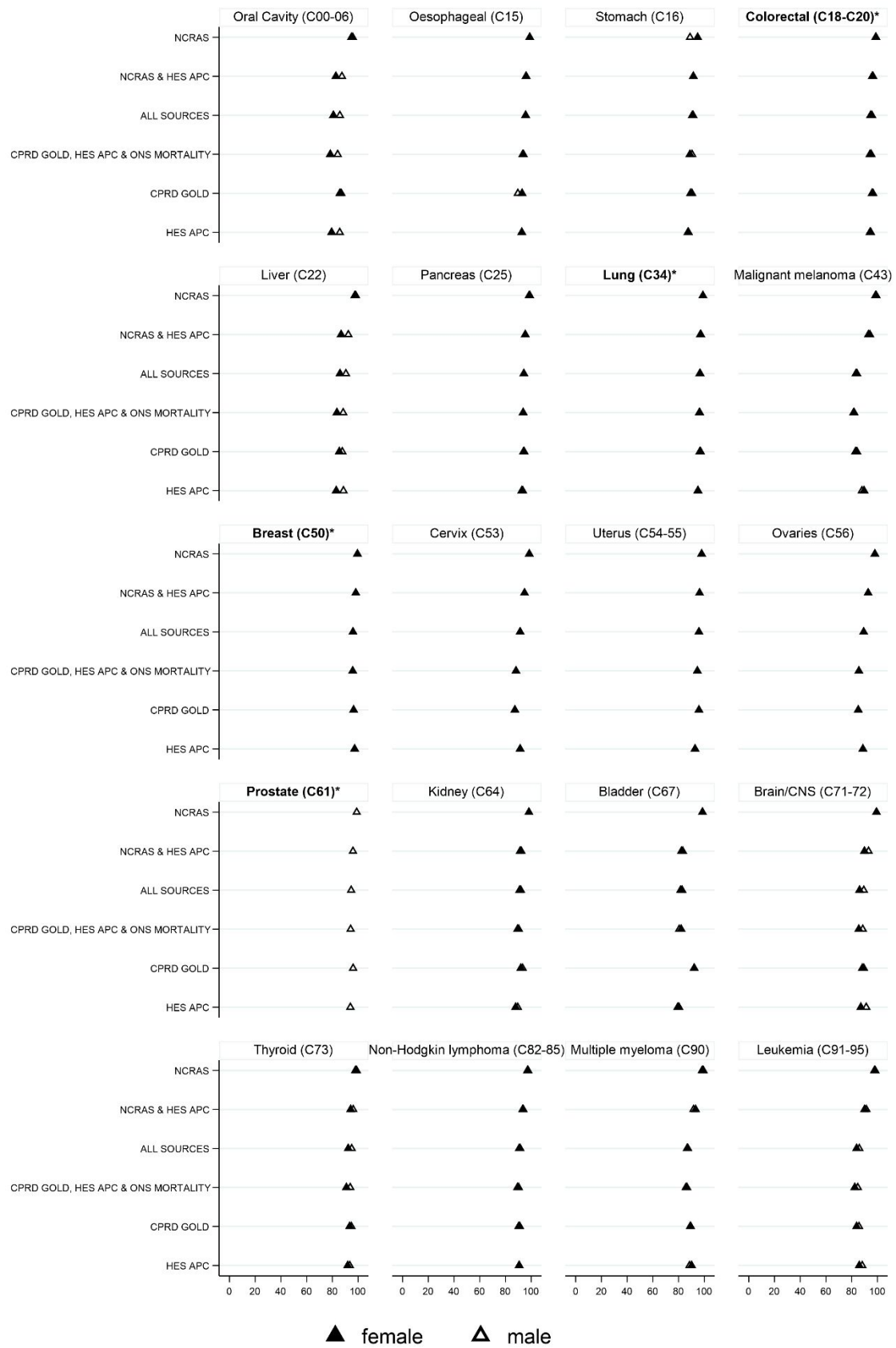


Figure 3: Positive Predictive Value by calendar year

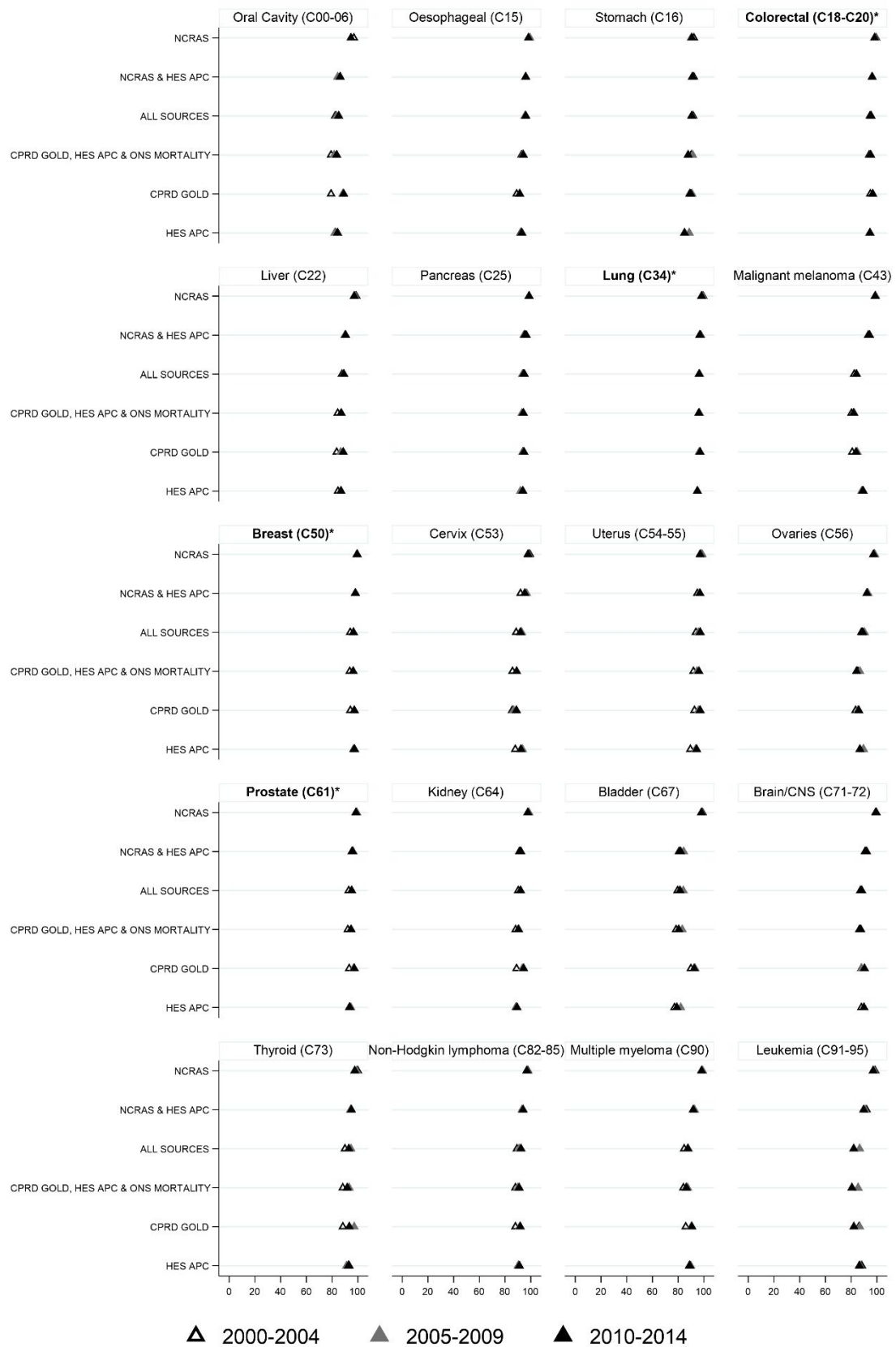


Figure 4: Sensitivity by age

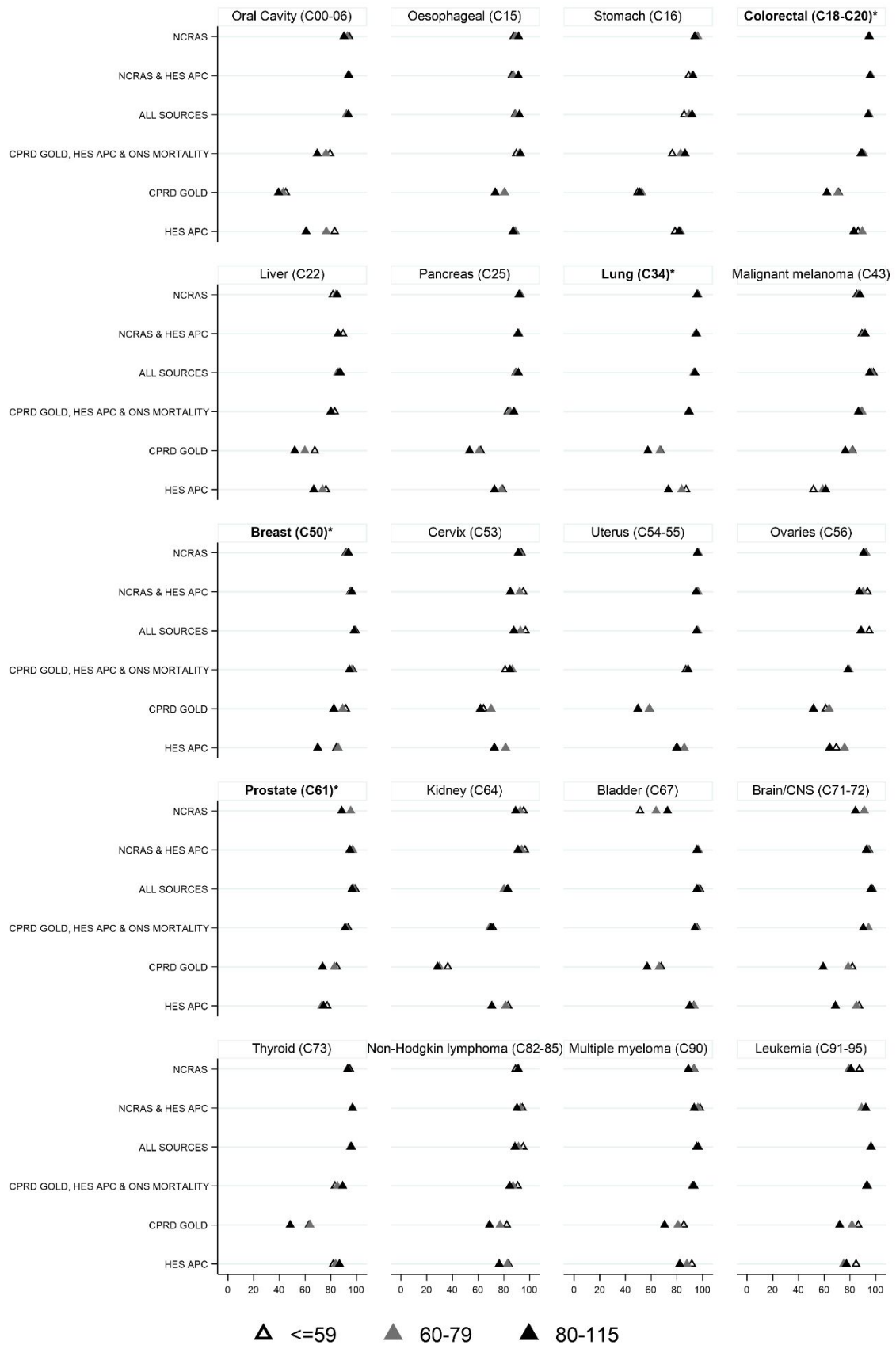




Figure 5: Sensitivity by sex

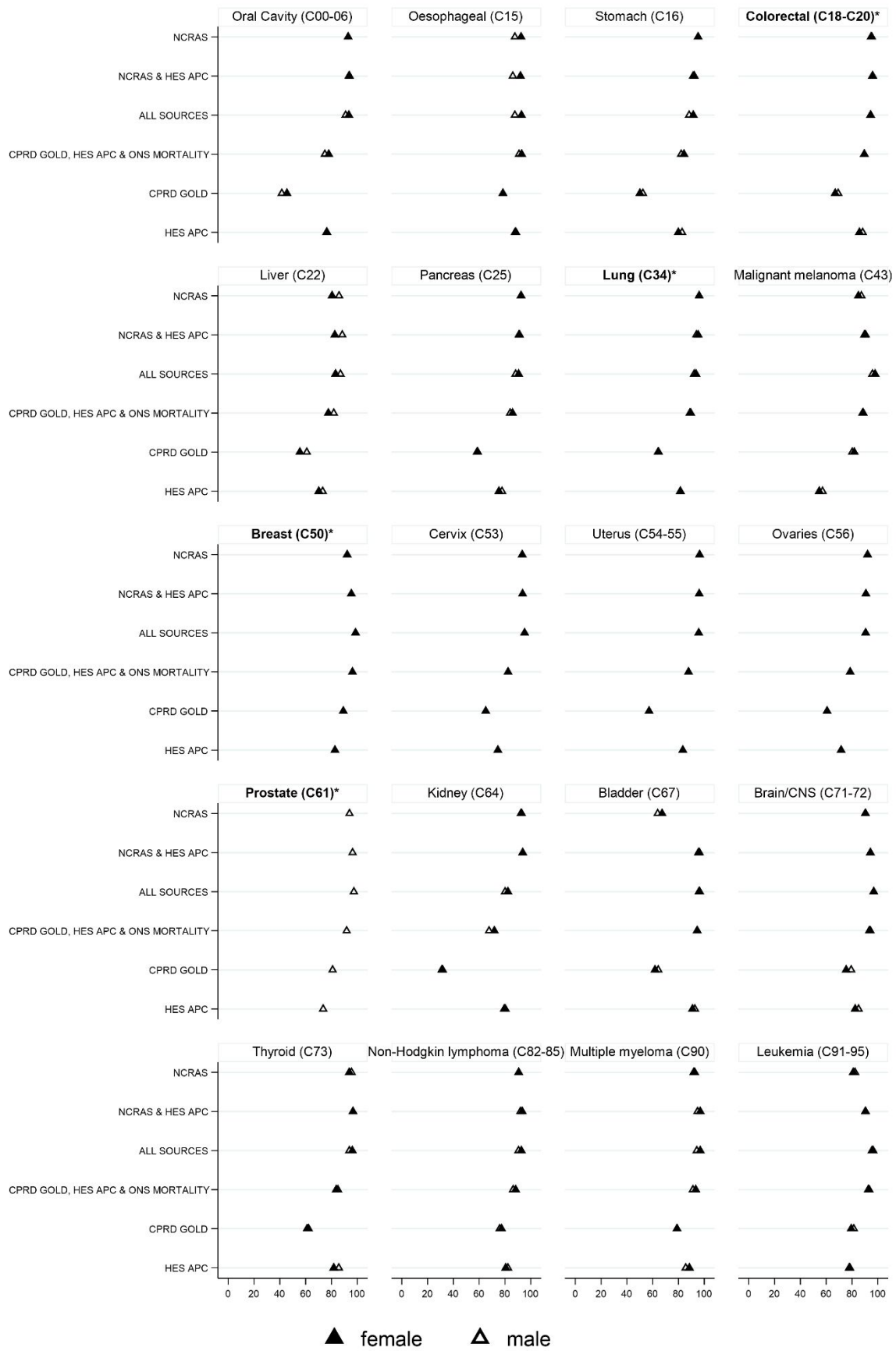


Figure 6: Sensitivity by calendar year

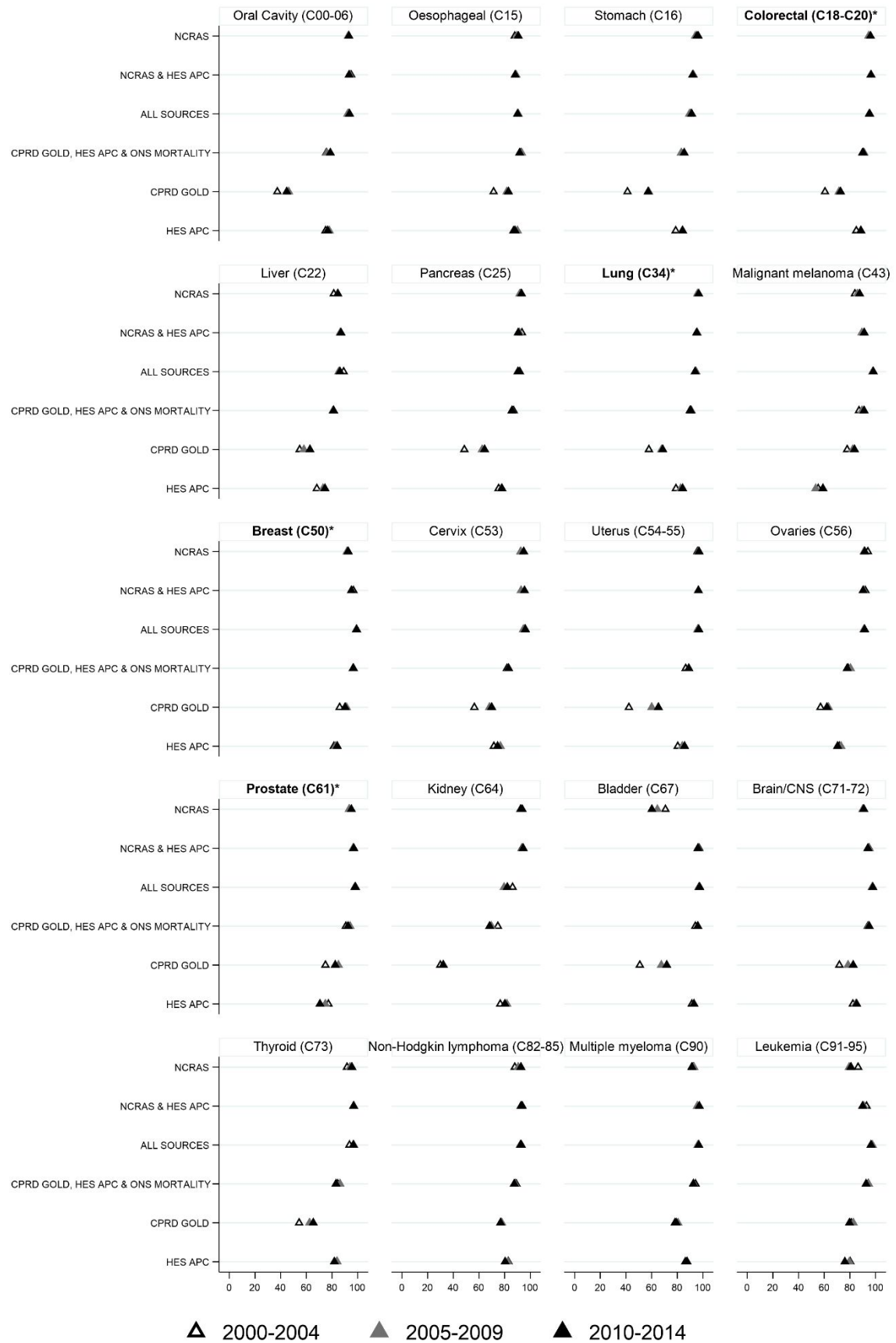


Figure 7: Output from logistic regression models with completeness of stage and grade as the dependent variables

Created using coefplot command in Stata <http://repec.sowi.unibe.ch/stata/coefplot/>

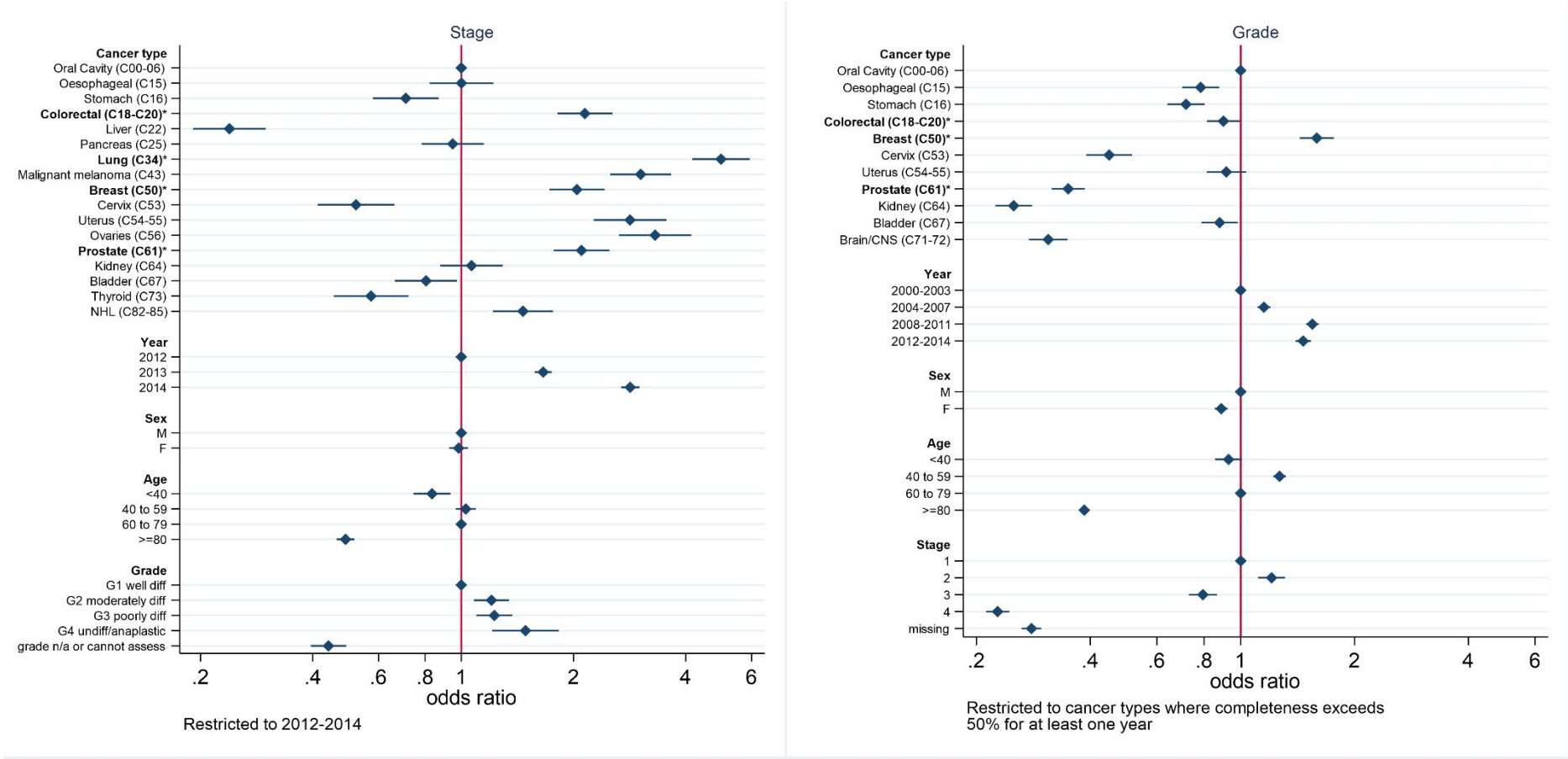
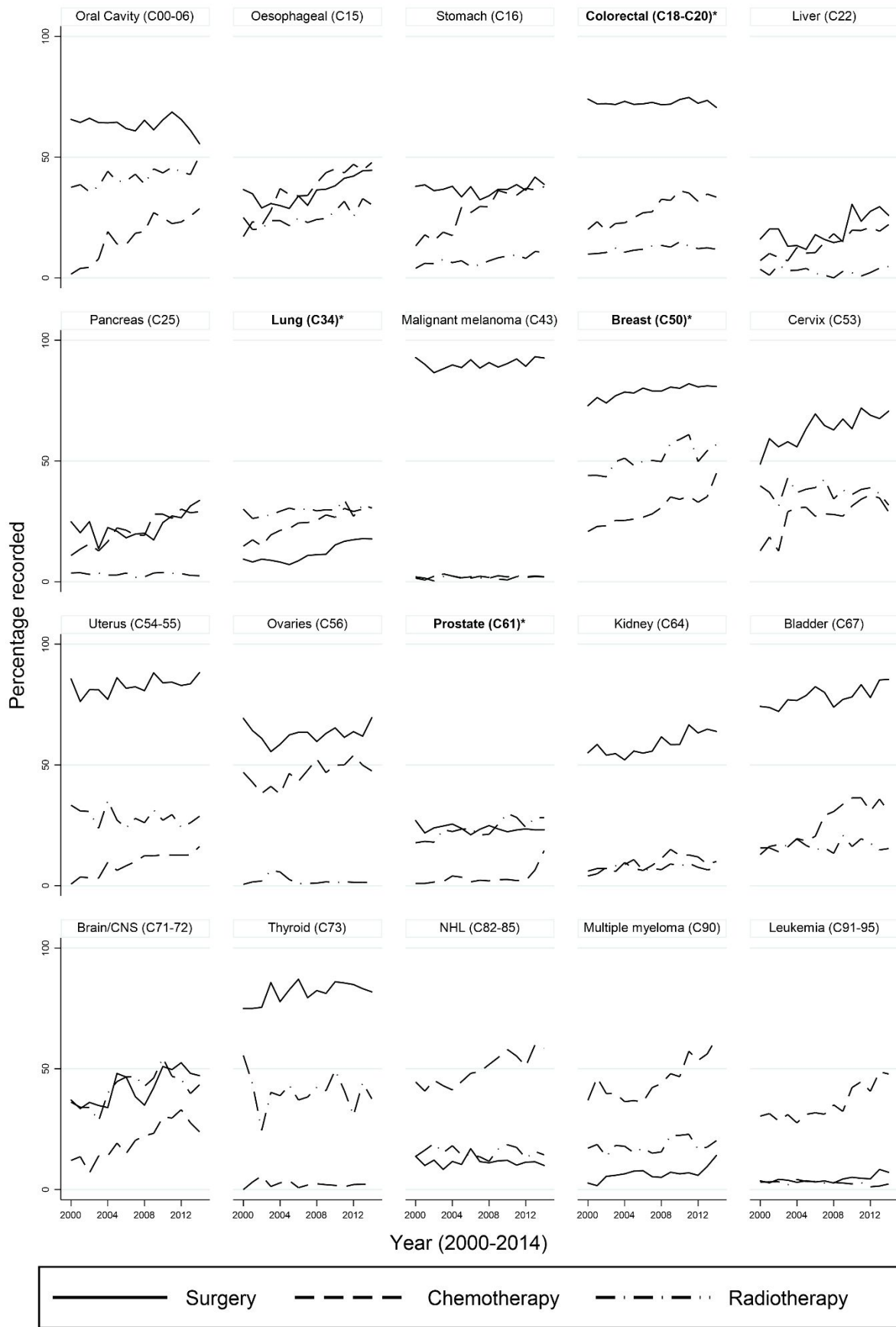


Figure 8: Recording of treatment modalities for patients identified using NCRAS data only



# BMJ Open

**What are the implications of using individual and combined sources of routinely collected data to identify and characterise incident site-specific cancers? A concordance and validation study using linked English electronic health records data**

Journal:	<i>BMJ Open</i>
Manuscript ID	bmjopen-2020-037719.R1
Article Type:	Original research
Date Submitted by the Author:	28-Apr-2020
Complete List of Authors:	Strongman, Helen; London School of Hygiene and Tropical Medicine, Department of Non-communicable Disease Epidemiology Williams, Rachael; Medicines and Healthcare Products Regulatory Agency, Clinical Practice Research Datalink (CPRD) Bhaskaran, Krishnan; London School of Hygiene & Tropical Medicine, Non-Communicable Disease Epidemiology
<b>Primary Subject Heading</b>:	Oncology
Secondary Subject Heading:	Epidemiology, Health informatics, Research methods
Keywords:	ONCOLOGY, EPIDEMIOLOGY, STATISTICS & RESEARCH METHODS, Health informatics < BIOTECHNOLOGY & BIOINFORMATICS

SCHOLARONE™  
Manuscripts



I, the Submitting Author has the right to grant and does grant on behalf of all authors of the Work (as defined in the below author licence), an exclusive licence and/or a non-exclusive licence for contributions from authors who are: i) UK Crown employees; ii) where BMJ has agreed a CC-BY licence shall apply, and/or iii) in accordance with the terms applicable for US Federal Government officers or employees acting as part of their official duties; on a worldwide, perpetual, irrevocable, royalty-free basis to BMJ Publishing Group Ltd ("BMJ") its licensees and where the relevant Journal is co-owned by BMJ to the co-owners of the Journal, to publish the Work in this journal and any other BMJ products and to exploit all rights, as set out in our [licence](#).

The Submitting Author accepts and understands that any supply made under these terms is made by BMJ to the Submitting Author unless you are acting as an employee on behalf of your employer or a postgraduate student of an affiliated institution which is paying any applicable article publishing charge ("APC") for Open Access articles. Where the Submitting Author wishes to make the Work available on an Open Access basis (and intends to pay the relevant APC), the terms of reuse of such Open Access shall be governed by a Creative Commons licence – details of these licences and which [Creative Commons](#) licence will apply to this Work are set out in our licence referred to above.

Other than as permitted in any relevant BMJ Author's Self Archiving Policies, I confirm this Work has not been accepted for publication elsewhere, is not being considered for publication elsewhere and does not duplicate material already published. I confirm all authors consent to publication of this Work and authorise the granting of this licence.

1  
2  
3 1 **What are the implications of using individual and combined sources of routinely collected data to**  
4  
5 2 **identify and characterise incident site-specific cancers? A concordance and validation study using**  
6  
7 3 **linked English electronic health records data.**  
8  
9

10 4 Authors: Helen Strongman PhD<sup>1</sup>, Rachael Williams PhD<sup>2</sup>, Prof Krishnan Bhaskaran<sup>1</sup>

11  
12  
13 5 <sup>1</sup> Department of Non-Communicable Diseases Epidemiology, London School of Hygiene and Tropical  
14 6 Medicine, London;

15  
16 7 <sup>2</sup> Clinical Practice Research Datalink, Medicines and Healthcare Products Regulatory Agency,  
17 8 London, UK  
19

20 9  
21  
22 10 Correspondence: Helen Strongman, Dept of Non-Communicable Disease Epidemiology, London  
23  
24 11 School of Hygiene and Tropical Medicine, London WC1E 7HT, [helen.strongman@lshtm.ac.uk](mailto:helen.strongman@lshtm.ac.uk),  
25  
26 12 +44(0)20 7636 8636  
27  
28  
29  
30  
31

32 14 Abstract: 300 words

33  
34  
35 15 Manuscript: 3352 words  
36  
37

38 16 Tables: 1  
39  
40

41 17 Figures: 5  
42  
43

44 18 Keywords: Cancer; Data Quality; Data Sources; Data Linkage; Epidemiologic Research Designs  
45  
46

## 47 19 **Abstract**

### 48 49 50 20 **Objectives**

51  
52  
53 21 To describe the benefits and limitations of using individual and combinations of linked English  
54  
55 22 electronic health data to identify incident cancers.  
56  
57

### 58 23 **Design and setting**

59  
60

1  
2  
3 24 Our descriptive study uses linked English Clinical Practice Research Datalink primary care; cancer  
4  
5 25 registration; hospitalisation and death registration data.  
6  
7

## 8 26 **Participants and measures**

9  
10  
11 27 We implemented case definitions to identify first site-specific cancers at the 20 most common sites,  
12  
13 28 based on the first ever cancer diagnosis recorded in each individual or commonly used combination  
14  
15 29 of data sources between 2000-2014.  
16  
17

18 30 We calculated positive predictive values and sensitivities of each definition, compared to a gold  
19  
20 31 standard algorithm that used information from all linked datasets to identify first cancers. We  
21  
22 32 described completeness of grade and stage information in the cancer registration dataset.  
23  
24  
25

## 26 33 **Results**

27  
28  
29 34 165 953 gold standard cancers were identified. Positive predictive values of all case definitions were  
30  
31 35  $\geq 80\%$  and  $\geq 94\%$  for the four most common cancers (breast, lung, colorectal, prostate).  
32  
33

34 36 Sensitivity for case definitions that used cancer registration alone or in combination was  $\geq 92\%$  for  
35  
36 37 the four most common cancers and  $\geq 80\%$  across all cancer sites except bladder cancer (65% using  
37  
38 38 cancer registration alone). For case definitions using linked primary care, hospitalisation and death  
39  
40 39 registration data, sensitivity was  $\geq 89\%$  for the four most common cancers, and  $\geq 80\%$  for all cancer  
41  
42 40 sites except kidney (69%), oral cavity (76%) and ovarian cancer (78%). When primary care or  
43  
44 41 hospitalisation data were used alone, sensitivities were generally lower and diagnosis dates were  
45  
46 42 delayed. Completeness of staging data in cancer registration data was high from 2012 (minimum  
47  
48 43 76.0% 2012 86.4% 2014 for the four most common cancers).  
49  
50  
51

## 52 44 **Conclusions**

53  
54  
55 45 Ascertainment of incident cancers was good when using cancer registration data alone or in  
56  
57 46 combination with other datasets, and for the majority of cancers when using a combination of  
58  
59 47 primary care, hospitalisation and death registration data.  
60



## 48 Article Summary

### 49 Strengths and limitations of the study

- 50 - This is the first study to present comprehensive information on the implications of using  
51 different individual and combinations of linked electronic health data sources in England to  
52 identify cases of the 20 most common incident cancers.
- 53 - Using a gold standard algorithm that combined all available data from multiple sources as a  
54 comparator, we were able to estimate both positive predictive values and sensitivity values  
55 for a range of pragmatic case definitions.
- 56 - We described similarities and differences in values between age groups, sexes and calendar  
57 years, the impact of choice of source(s) on diagnosis dates and mortality rates, and  
58 completeness of stage and grade in cancer registration data.
- 59 - A key limitation was that our gold standard algorithm is not validated and may be affected  
60 by differences in clinical diagnosis and coding of invasive cancers between data sources.

## 61 Introduction

62 The Clinical Practice Research Datalink provides de-identified primary care data linked to additional  
63 secondary health data sources, under a well-governed framework<sup>1</sup>. Use of linked data helps  
64 researchers to answer more epidemiological questions and increase study quality through improved  
65 exposure, outcome and covariate classification<sup>2</sup>. In the field of cancer epidemiology, CPRD primary  
66 care data linked to Hospital Episode Statistics Admitted Patient Care data (HES APC), Office of  
67 National Statistics (ONS) mortality, and National Cancer Registration and Analysis Service (NCRAS)  
68 cancer registration data are used to analyse factors contributing to the risk of cancer and the  
69 consequences of cancer and its treatment. Use of linked data reduces the sample to the common  
70 source population and data coverage period for each included dataset, and has cost and logistical  
71 implications, which are greatest for NCRAS data. Research teams therefore commonly choose not to

1  
2  
3 72 use all available linked data<sup>3</sup>. Cancer epidemiology studies can also be conducted using NCRAS and  
4  
5 73 HES APC data provided by NHS Digital and Public Health England (PHE), without linkage to CPRD  
6  
7 74 primary care data<sup>4</sup>. This provides national coverage at the expense of the detailed health data that  
8  
9  
10 75 are available in primary care records.

11  
12  
13 76 Validation studies assessing concordance between CPRD GOLD, HES APC and NCRAS data have  
14  
15 77 estimated high Positive Predictive Values (PPVs) for CPRD GOLD data and varying proportions of  
16  
17 78 registered cancers that are not captured in CPRD GOLD and HES APC<sup>5-8</sup>. The most up to date analysis  
18  
19 79 by Arhi et al. included the 5 most common cancers and all papers focused on concordance between  
20  
21  
22 80 CPRD GOLD only and NCRAS; existing evidence therefore does not provide a complete assessment of  
23  
24 81 the benefits and limitations of using different combinations of data sources within the context of  
25  
26 82 practical study designs. National data are available describing completeness of data fields within the  
27  
28 83 cancer registry data in each collection year<sup>9</sup> and over time for all cancers combined<sup>4</sup>; missingness for  
29  
30 84 individual years has been associated with age, comorbidities and Clinical Commissioning Groups<sup>10,11</sup>.

31  
32  
33 85 We aim to describe and compare the benefits and limitations of using different combinations of  
34  
35 86 linked CPRD primary care data, HES APC, ONS mortality, and NCRAS cancer registration data, for  
36  
37 87 conducting cancer epidemiology studies. Our analyses focus on incident cancer ascertainment as it is  
38  
39 88 a common and important outcome in cancer epidemiology, and it is more difficult to distinguish  
40  
41 89 between secondary, recurrent and primary cancers at a second site in these datasets. We have  
42  
43 90 compared definitions of the twenty most common cancers based on the first ever cancer recorded in  
44  
45 91 individual or combinations of datasets with a gold standard definition comparing information from  
46  
47 92 all four datasets. We also describe the availability of stage, grade and treatment variables over time  
48  
49 93 in the cancer registration data for the CPRD linked cohort. This reflects real life study design and will  
50  
51 94 help researchers to decide which combination of data sources to use for future studies.  
52  
53  
54  
55

56 95

## 57 96 **Methods**

## 97 **Study design and setting**

98 We completed a concordance study using linked<sup>2</sup> English CPRD GOLD, HES APC, ONS mortality and  
99 NCRAS data. CPRD GOLD data were extracted from the January 2017 monthly release and the 13<sup>th</sup>  
100 update to CPRD's linked data. The study period was 1 Jan 2000 – 31 December 2014, with 31  
101 December matching the end of the NCRAS coverage period.

102 The CPRD GOLD database includes de-identified records from participating general practices in the  
103 United Kingdom (UK) who use Vision software<sup>1</sup>. General practice staff can record cancer diagnoses  
104 using Read codes or in free text comments boxes, though the latter are not collected by CPRD.  
105 Diagnoses will typically be entered during/following a consultation or from written information that  
106 is returned to the practice from secondary care. CPRD GOLD data are linked to HES APC, ONS  
107 mortality and NCRAS through a trusted third party for English practices that have agreed to  
108 participate in the linkage programme<sup>2</sup>. HES APC data are collected by NHS Digital to co-ordinate  
109 clinical care in England and calculate hospital payments<sup>12</sup>. Admissions for and related to cancer  
110 diagnoses are recorded using ICD-10 codes. National cancer registration data are collected by NCRAS  
111 which is part of Public Health England (PHE) in accordance with the Cancer Outcomes and Services  
112 Dataset (COSD)<sup>13</sup> which has been the national standard for reporting of cancer in England since  
113 January 2013. Data include ICD-10 codes to identify the cancer site and more detailed information  
114 such as stage and grade. ONS mortality data includes dates and causes of deaths registered in  
115 England, recorded using ICD-10 codes.

## 116 **Participants, exposures and outcomes**

117 Our underlying study population included male and female patients registered in CPRD GOLD  
118 practices who were eligible for linkage to HES APC, NCRAS and ONS mortality data and had at least  
119 366 days of follow-up between 1 January 1999 and 31 December 2014. Start of follow-up was  
120 defined as the latest of the current registration date within the practice and the CPRD estimated  
121 start of continuous data collection for the practice (up-to-standard date). End of follow-up was

1  
2  
3 122 determined as the date the patient left the practice, ONS mortality date of death, or practice last  
4  
5 123 collection date.

6  
7 124

8  
9  
10 125 *Identification and classification of cancer codes:* We used code lists to classify cancer records in each  
11  
12 126 of CPRD GOLD, HES APC, and ONS mortality data as one of the 20 most common sites, other  
13  
14 127 specified cancers, history of cancer, secondary cancers, benign tumours, administrative cancer  
15  
16 128 codes, unspecified and incompletely specified cancer codes

17  
18  
19 129 (<https://doi.org/10.17037/data.00001519>). Incompletely specified cancer codes could be mapped to  
20  
21 130 >1 cancer site (e.g. ICD10 code C68.9 “Malignant neoplasms of urinary organ unspecified” was  
22  
23 131 considered consistent with both bladder and kidney cancer). For NCRAS, we accessed coded records  
24  
25 132 for the 20 most common cancers. We included cancers recorded in the clinical or referral file for  
26  
27 133 CPRD GOLD, cancers recorded in any diagnosis field for HES APC, and the underlying or most  
28  
29 134 immediate cancer cause of death in ONS mortality data.

30  
31  
32  
33 135 *Cancer case definitions based on individual sources and combinations of sources:* We developed  
34  
35 136 alternative cancer case definitions mirroring those commonly used in epidemiology studies, based  
36  
37 137 on identifying the first malignant cancer (excluding administrative codes and benign tumours)  
38  
39 138 recorded in various combinations of data sources (NCRAS alone; NCRAS and HES APC; all sources;  
40  
41 139 CPRD GOLD, HES APC and ONS mortality; CPRD GOLD alone, HES APC alone). Multiple malignant  
42  
43 140 cancers recorded on the index date in CPRD GOLD or HES APC were reclassified as multiple-site  
44  
45 141 cancer and were not considered as individual-site cancer records for positive predictive value and  
46  
47 142 sensitivity calculations; multiple codes recorded in different sources on the same date were  
48  
49  
50 143 reclassified as the site identified in the NCRAS data if available and as multiple-site cancer if not. For  
51  
52 144 each case definition, we only examined the first malignant cancer per individual where this occurred  
53  
54  
55 145 within the study period and at least one year after the start of follow-up.

1  
2  
3 146 *Gold standard cancer case definition:* We developed a gold standard algorithm that classifies  
4  
5 147 incident records of the 20 most common cancers by comparing the first malignant cancer identified  
6  
7 148 in each individual source (Figure 1). Cancers recorded in NCRAS alone with no contradictions (i.e.  
8  
9 149 records for first cancers at different sites) were considered true cases whereas cancers recorded in  
10  
11 150 HES APC alone or GOLD alone required internal confirmation within that source in the form of  
12  
13 151 another code for cancer consistent with the same site (or with site unspecified) within 6 months and  
14  
15 152 no contradictory codes (e.g. for cancers at other sites) in this period. Where cancer records were  
16  
17 153 present in >1 data source, we considered a site-specific cancer to be a true case (a) if it was recorded  
18  
19 154 as the first cancer in NCRAS and the total number of data sources with records for cancer at that site  
20  
21 155 was equal to or greater than the number of data sources with contradictory records (i.e. records for  
22  
23 156 first cancers at different sites); or (b) where the cancer was not present in NCRAS, if there were  
24  
25 157 more data sources in total with records for cancer at that site than data sources with contradictory  
26  
27 158 records.

28  
29  
30  
31  
32  
33 159 We used NCRAS data to identify stage, grade and treatment where available in the cancer registry  
34  
35 160 only cohort. Binary surgery, chemotherapy and radiotherapy variables were derived using individual  
36  
37 161 records of treatment from the first year after diagnosis.

#### 38 39 40 162 **Statistical analysis**

41  
42  
43 163 For each cancer site and each individual or combined data source, we combined our applied study  
44  
45 164 definitions with our gold standard definition to classify each applied study definition as a true  
46  
47 165 positive, false positive, or false negative record.

48  
49  
50 166 We used these categories to calculate sensitivity and positive predictive value overall and stratified  
51  
52 167 by age categories (<60, 60-79, 80+), calendar year and sex. We calculated differences in diagnosis  
53  
54 168 dates for true positives by subtracting the gold standard index date from the index date for each  
55  
56 169 source and combination of sources.  
57  
58  
59  
60

1  
2  
3 170 We used Kaplan-Meier methods to describe mortality over time for cancers identified using each  
4  
5 171 definition. The ONS mortality death date was used for these analyses.  
6  
7

8 172 We used the NCRAS only definition to calculate proportions of patients with complete stage and  
9  
10 173 grade and recorded cancer treatment modalities over time.  
11  
12

### 13 174 **Patient public involvement**

14  
15  
16 175 Patients and the public were not involved in conceiving, designing or conducting this study and will  
17  
18 176 not be consulted regarding the dissemination of study results.  
19  
20

21 177 This study was approved by the London School of Hygiene & Tropical Medicine Ethics Committee  
22  
23 178 (6202) and the Independent Scientific Advisory Committee for the Medicines and Healthcare  
24  
25 179 products Regulatory Agency database research (12\_068R).  
26  
27

### 28 29 180 **Results**

30  
31  
32 181 Of 14 747 047 research quality patients in the CPRD GOLD January 2017 build, 8 893 326 were  
33  
34 182 eligible for linkage to HES, ONS mortality and NCRAS data in set 13; 237 were excluded due to  
35  
36 183 unknown sex. Of the remainder, 6 791 074 had at least one year of follow-up between 1 January  
37  
38 184 1999 and 31 December 2014 and were included in the study population. Using the gold standard  
39  
40 185 algorithm, 165 953 incident cases of cancer were identified. The number of patients identified with  
41  
42 186 each cancer is presented in supplementary appendix table 1. Half (50.0%, n=82 899) of these  
43  
44 187 patients were male; 24.4% (40 470) aged 0-59, 54.1% (89 720) aged 60-79 and 21.6% (35 763) aged  
45  
46 188 80 or older.  
47  
48  
49

50  
51 189 Figure 2 presents PPVs for each case definition, comparing the first recorded cancer in each  
52  
53 190 combination of data sources with the gold standard algorithm. When using NCRAS data alone, 91.0%  
54  
55 191 to 99.5% of cancers were confirmed by the algorithm; for 19 out of 20 cancer sites, the NCRAS-only  
56  
57 192 case definition gave the highest PPV. Case definitions using data sources not including NCRAS  
58  
59 193 generally had lower PPVs, ranging from 79.6% to 97.3% for individual cancer sites. For the four most  
60

1  
2  
3 194 common cancers (breast, lung, colorectal, prostate), PPVs were at least 94% for all case definitions.  
4  
5 195 Minimal differences in PPVs were observed between age groups, years and sexes (supplementary  
6  
7 196 appendix figures 1 to 3).  
8  
9  
10 197 Figure 3 presents sensitivity values for each case definition. Sensitivity was generally higher for the  
11  
12 198 case definitions that included NCRAS data (ranging from 80.9 to 98.7% for individual cancer sites  
13  
14 199 except bladder cancer identified using NCRAS data alone [64.8%], and  $\geq 92\%$  for the four most  
15  
16 200 common cancers [breast, lung, colorectal, prostate]). Sensitivity was also generally high for  
17  
18 201 definitions using a combination of CPRD GOLD, HES APC and ONS mortality data (ranging from 69.2  
19  
20 202 to 96.3%,  $\geq 89\%$  for the four most common cancers). Sensitivity was lower for case definitions that  
21  
22 203 used CPRD GOLD alone (range 31.5-89.3% for individual cancer sites) or HES APC alone (range 55.9-  
23  
24 204 92.6%). Sensitivity values for CPRD GOLD alone and HES APC alone increased slightly in younger  
25  
26 205 patients and more recent years; no differences were observed between males and females  
27  
28 206 (supplementary appendix figures 4 to 6). Post-hoc analysis suggested that the low sensitivity of CPRD  
29  
30 207 GOLD only definitions for kidney cancer (sensitivity 31.5%, n false negatives 2869) was driven by  
31  
32 208 missing (n = 1 136, 39.6%) or incompletely specified urinary organ cancer codes (n = 1 108, 38.6%) in  
33  
34 209 CPRD GOLD rather than contradictory information about the first cancer record (n = 625, 21.8%).  
35  
36 210 These incompletely specified codes are less likely to be used for bladder cancers (n=85) than kidney  
37  
38 211 cancers (n=1 108). Bladder cancers that were not recorded in NCRAS data (n=3 445) were commonly  
39  
40 212 recorded in both HES APC and CPRD GOLD (n=2 228, 64.7%) or in HES APC only with a subsequent  
41  
42 213 unspecified or bladder cancer record in HES APC within 6 months (n=995, 28.9%).  
43  
44  
45  
46  
47  
48  
49 214 Table 1 describes the number of days (median IQR and 5<sup>th</sup>/95<sup>th</sup> percentile) lag between the date of  
50  
51 215 incident cancers from the gold standard definition and the date of cancer arising from each case  
52  
53 216 definition (i.e. the first record within the specific combinations of data sources used). Case  
54  
55 217 definitions using NCRAS alone and combinations of  $\geq 2$  data sources captured cancers close to the  
56  
57  
58  
59  
60

1  
2  
3 218 gold standard date (median lag  $\leq 7$  days for all cancer sites), whereas median lags were generally  
4  
5 219 longer for the case definitions using CPRD GOLD alone and HES APC alone.  
6  
7  
8 220 Figure 4 describes mortality over time following incident cancer diagnoses ascertained from each  
9  
10 221 case definition. Minimal differences in mortality were observed between cancers identified from  
11  
12 222 different case definitions. Where variability was observed, cancers identified using CPRD GOLD only  
13  
14 223 had the lowest mortality rates (e.g. kidney cancer) and cancers identified using HES APC only or  
15  
16 224 NCRAS only had higher mortality rates (e.g. prostate cancer and bladder cancer respectively).  
17  
18  
19  
20 225 Figure 5 describes completeness of grade and stage for cancers identified using NCRAS only.  
21  
22 226 Recording of grade was highly variable between cancers with gradual increases in completeness over  
23  
24 227 time. Completeness of staging information was low in earlier calendar years but improved  
25  
26 228 substantially from around 2012 especially for the four most common cancers (minimum 76.0% 2012,  
27  
28 229 86.4% 2014). Post-hoc logistic regression models adjusted for year and cancer site indicated that  
29  
30 230 completeness of stage and grade were associated with each other and these variables were least  
31  
32 231 complete in patients aged  $\geq 80$ ; stage data was more complete for higher grade tumours whereas  
33  
34 232 grade data was more complete for lower stage tumours (supplementary appendix figure 7).  
35  
36  
37  
38 233 Supplementary appendix figure 8 describes recording of treatment modalities identified using  
39  
40 234 NCRAS only. Missing records may indicate that the patient did not receive that treatment modality  
41  
42 235 or that the treatment modality was not recorded.  
43  
44  
45

## 236 Discussion

### 237 *Statement of principal findings*

238 We investigated the use of different sources of electronic health record data to identify incident  
239 cancers. For all case definitions, using individual or combined data sources, a minimum of 80% of  
240 incident site-specific cancers were confirmed using the gold standard algorithm; this rose to 94% of  
241 the four most common cancers. Use of cancer registration data alone or in any combination of data



1  
2  
3 242 sources captured at least 80% of site-specific cancers identified by the gold standard algorithm,  
4  
5 243 excepting bladder cancer, and 92% of cases for the four most common cancers. Combining CPRD  
6  
7 244 GOLD, HES APC and ONS mortality data captured at least 80% of site-specific cancers excepting  
8  
9 245 kidney, oral cavity and ovarian cancers, and captured  $\geq 89\%$  of cases for the four most common  
10  
11 246 cancers. Sensitivity was much more variable when using primary care or hospital data alone, and  
12  
13 247 dropped to 65% when identifying bladder cancers using cancer registration data alone. Use of  
14  
15 248 primary care or hospital data alone resulted in a small lag in identifying cancers of interest,  
16  
17 249 compared to the gold standard dates but other case definitions captured cancers close to the gold  
18  
19 250 standard date. Finally, whilst we observed minimal changes in PPVs and sensitivities between 2000  
20  
21 251 and 2014, completeness of NCRAS cancer registration stage and grade data increased markedly  
22  
23 252 from 2012 onwards for specific cancer types, demonstrating that initiatives to improve data can  
24  
25 253 have a profound impact on the quality of the data<sup>4</sup>. Completeness of cancer treatment recording  
26  
27 254 was difficult to assess due to the absence of a missing category.  
28  
29  
30  
31  
32  
33  
34  
35

### 36 256 *Strengths and weaknesses of the study*

37  
38 257 The main strength of this study is that we have developed a gold standard algorithm using the  
39  
40 258 entirety of the evidence available from CPRD to demonstrate the impact of choice of datasets in  
41  
42 259 identifying incident cancers for real life studies. We have also assessed the value of using NCRAS  
43  
44 260 cancer registration data to measure stage, grade and cancer treatment modalities.  
45  
46  
47  
48 261 A limitation of the study is that our gold standard algorithm is not validated. We feel that we were  
49  
50 262 justified in pre-weighting NCRAS data as more reliable than other data sources as NCRAS is a highly  
51  
52 263 validated data set that matches, merges and quality checks data from multiple sources<sup>4</sup>. We did not  
53  
54 264 consider NCRAS to be the outright gold standard as it is plausible that NCRAS does not identify all  
55  
56 265 tumours diagnosed and treated in primary and secondary care. For most cancer sites, our gold  
57  
58 266 standard algorithm identified a small proportion of cancers that are recorded in HES APC, CPRD  
59  
60

1  
2  
3 267 GOLD or ONS mortality data but not in NCRAS. These tumours may have been diagnosed and coded  
4  
5 268 as invasive in primary or secondary care but not by NCRAS; been incorrectly coded in HES APC, CPRD  
6  
7 269 GOLD or ONS mortality data; not have been notified to NCRAS (e.g. tumours treated in private  
8  
9  
10 270 hospitals); or be the result of linkage errors between the data sets. The proportion of cancers  
11  
12 271 identified in HES APC but not in NCRAS is particularly high for bladder cancer. This is likely to be the  
13  
14 272 result of difficulties, inconsistencies and changes in the pathological definition and coding of cancers  
15  
16 273 over time in NCRAS, which are greatest for bladder cancer<sup>4,14</sup>. This explanation is supported by the  
17  
18 274 higher mortality rates that we observed in bladder cancer cases identified in NCRAS compared with  
19  
20 275 other data sources. To identify incident cancers, we required 12 months of research quality follow-  
21  
22 276 up in CPRD GOLD prior to inclusion in the study. Previous research has demonstrated that historic  
23  
24 277 data is generally incorporated within the patient record with this time frame<sup>15</sup> The identification of  
25  
26 278 first ever cancers will also have been affected by different lengths of follow-up data available in  
27  
28 279 linked data sources as NCRAS data collection started in 1990, HES APC in 1997 and ONS mortality  
29  
30 280 data in 1998, and by the inclusion of all diagnostic codes in HES APC assuming that the first ever  
31  
32 281 primary or secondary record identified incident cancer. Reassuringly, PPVs for liver and brain cancer  
33  
34 282 were high for all individual and combinations of datasets suggesting that these were not unduly  
35  
36 283 misclassified as primary incident cancers despite being common sites for metastases. Requiring  
37  
38 284 internal confirmation within 6 months for cancers recorded in CPRD GOLD alone in our GOLD  
39  
40 285 standard definition is more likely to discount cancers with poorer prognoses and those recorded in  
41  
42 286 the last 6 months of follow-up. Our data cut only included NCRAS data for the top 20 cancers; earlier  
43  
44 287 cancers at other sites will have been missed in this study.  
45  
46  
47  
48  
49  
50 288 It is also important to note that as the gold standard algorithm uses data recorded after the first  
51  
52 289 record of the cancer site in any source (index date), it cannot be used to identify outcomes in applied  
53  
54 290 studies and follow-up of cohort studies with cancer as an exposure would need to start at least 6  
55  
56 291 months after diagnosis; our first ever cancer record in any source definition would be more  
57  
58 292 appropriate for most studies.  
59  
60

1  
2  
3 2934  
5 294 *Strengths and weaknesses in relation to other studies, discussing important differences in results*6  
7  
8 295 The most up to date study describing concordance between linked CPRD GOLD, HES APC and NCRAS9  
10 296 datasets demonstrated that 2-4% of the 5 most common cancers recorded in CPRD GOLD are not11  
12 297 confirmed in either HES APC or cancer registration data and 9-33% of registered cancers are not13  
14 298 recorded in CPRD GOLD<sup>8</sup>. For cancers recorded in both sources, the diagnosis date was a median of15  
16 299 6-16 days later in CPRD GOLD than in the registration data. Using CPRD GOLD alone to identify these17  
18 300 cancers marginally over represented younger, healthier patients and identified 1-6% fewer deaths in19  
20 301 the first five years after diagnosis. Use of HES APC only identified a higher proportion of patients21  
22 302 with the correct diagnosis date than CPRD GOLD, but over represented older patients and those23  
24 303 diagnosed through the emergency route. The majority of registered cancers were picked up using25  
26 304 both CPRD GOLD and HES APC (ranging from 91% for lung cancer to 97% for breast cancer). Previous27  
28 305 research demonstrated similar results with substantial differences between cancer types<sup>5,6</sup>.29  
30 306 Additionally, a study using data from 2001-2007 found that using HES data in addition to NCRAS data31  
32 307 identified an additional 1.9%, 0.4% and 2.0% of surgically treated colorectal, lung and breast cancer33  
34 308 cases respectively<sup>16</sup>.35  
36 309 Our study is consistent with these results and provides more complete and practical evidence of the37  
38 310 strengths and limitations of using individual and combinations of linked datasets to identify and39  
40 311 characterise the twenty most common incident cancers.41  
42 312 We have also demonstrated the added value of using cancer registration data to measure stage and43  
44 313 grade of incident cancers from about 2012 onwards. Levels of data completeness of staging45  
46 314 information in the CPRD extract in 2012 were similar to those reported by the United Kingdom and47  
48 315 Ireland Association of Cancer Registries (UKAICR)<sup>9</sup>.49  
50 31651  
52 317 *Meaning of the study: possible explanations and implications for clinicians and policymakers*53  
54  
55  
56  
57  
58  
59  
60

1  
2  
3 318 Use of NCRAS cancer registration data maximised the proportion of cases confirmed as true positive  
4  
5 319 based on all available linked information and captured the highest proportion of true positive cases;  
6  
7 320 highly complete staging and grading information is available from this source from approximately  
8  
9 321 2012. Case definitions based on a combination of CPRD GOLD, HES APC and ONS mortality data also  
10  
11 322 had acceptable validity for the majority of cancer sites including the four most common cancers.  
12  
13  
14  
15 323 These findings should be considered when deciding which data sources to include in research studies  
16  
17 324 and which sources to use to define cancer exposures, outcomes and covariates.  
18  
19  
20  
21 325

### 22 326 *Unanswered questions and future research*

23  
24  
25 327 Further research is required to investigate the validity of cancer recorded in CPRD GOLD and HES  
26  
27 328 APC that are not recorded in the NCRAS data and to understand differences in cancer data recording  
28  
29 329 with CPRD GOLD and CPRD Aurum, CPRD's recently launched primary care database based on  
30  
31 330 records from practices that use EMIS software<sup>17</sup>. Further investigation would be required to  
32  
33 331 confidently identify subtypes of cancer, either using codes available in each dataset (e.g. colon and  
34  
35 332 rectal cancer) or additional information available in HES APC or NCRAS data. Use of NCRAS's recently  
36  
37 333 launched Systemic Anti-Cancer Therapy (SACT)<sup>18</sup> and National Radiotherapy Datasets will also  
38  
39 334 improve ascertainment of therapies for future studies.  
40  
41  
42  
43

### 44 335 *Conclusion*

45  
46 336 Completeness and accuracy of recording of cancers in English data sources is high particularly when  
47  
48 337 using NCRAS cancer registration data alone or in any combination with other data sources, and for  
49  
50 338 the majority of cancers when using a combination of CPRD GOLD, HES APC and ONS mortality data.  
51  
52 339 Completeness of cancer stage and grade variables in NCRAS was low before 2012 but appears to  
53  
54 340 have substantially improved for most cancers in more recent calendar periods. It is not possible to  
55  
56 341 validate completeness of the available treatment data; these should be used with caution. This study  
57  
58 342 describes likely levels of misclassification for a range of data sources, combinations and cancer sites  
59  
60

1  
2  
3 343 enabling cancer epidemiologists to optimise study design and better understand the limitations of  
4  
5 344 their research.  
6  
7

### 8 345 **Funding**

10  
11 346 CPRD funded access to the linked data sources used in this work. This work was additionally  
12  
13 347 supported by the Wellcome Trust and Royal Society grant number 107731/Z/15/Z.  
14  
15

### 16 348 **Acknowledgements**

17  
18  
19  
20 349 This study is based in part on data from the Clinical Practice Research Datalink obtained under  
21  
22 350 licence from the UK Medicines and Healthcare products Regulatory Agency. The data is provided by  
23  
24 351 patients and collected by the NHS as part of their care and support. The interpretation and  
25  
26 352 conclusions contained in this study are those of the author/s alone.  
27  
28

### 29 353 **Protocol**

30  
31  
32  
33 354 Available on request  
34  
35

### 36 355 **Competing Interests**

37  
38  
39 356 RW is employed by CPRD. HS and KB have academic honorary contracts at PHE for a separate  
40  
41 357 collaborative research study.  
42  
43

### 44 358 **Contributions**

45  
46  
47 359 HS, RW and KB conceived the study and contributed to the study design. HS and KB did the data  
48  
49 360 management. HS did the statistical analysis and wrote the first draft. HS, RW and KB contributed to  
50  
51 361 subsequent drafts.  
52  
53  
54

### 55 362 **Patient consent for publication**

56  
57  
58 363 Not required  
59  
60

1  
2  
3 364 **Data sharing**  
4  
5

6 365 Data were obtained from the Clinical Practice Research Datalink, provided by the UK Medicines and  
7  
8 366 Healthcare products Regulatory Agency. The authors' licence for using these data does not allow  
9  
10 367 sharing of raw data with third parties. Information about access to Clinical Practice Research  
11  
12 368 Datalink data is available here: <https://www.cprd.com/research-applications>. Code lists for this  
13  
14 369 study are available at <https://doi.org/10.17037/data.00001519>  
15  
16  
17

18 370 **References**  
19  
20

- 21 371 1 Herrett E, Gallagher AM, Bhaskaran K, *et al*. Data Resource Profile: Clinical Practice Research  
22  
23 372 Datalink (CPRD). *Int J Epidemiol* 2015; **44**: 827–36.  
24  
25  
26 373 2 Padmanabhan S, Carty L, Cameron E, Ghosh RE, Williams R, Strongman H. Approach to record  
27  
28 374 linkage of primary care data from Clinical Practice Research Datalink to other health-related  
29  
30 375 patient data: overview and implications. *Eur J Epidemiol* 2019; **34**: 91–9.  
31  
32  
33 376 3 Badrick E, Renehan I, Renehan AG. Linkage of the UK Clinical Practice Research Datalink with  
34  
35 377 the national cancer registry. *Eur J Epidemiol* 2019; **34**: 101–2.  
36  
37  
38 378 4 Henson KE, Elliss-Brookes L, Coupland VH, *et al*. Data Resource Profile: National Cancer  
39  
40 379 Registration Dataset in England. *Int J Epidemiol* 2020; **49**: 16-16h.  
41  
42  
43 380 5 Dregan A, Moller H, Murray-Thomas T, Gulliford MC. Validity of cancer diagnosis in a primary  
44  
45 381 care database compared with linked cancer registrations in England. Population-based cohort  
46  
47 382 study. *Cancer Epidemiol* 2012; **36**: 425–9.  
48  
49  
50  
51 383 6 Boggon R, Van Staa TP, Chapman M, Gallagher AM, Hammad TA, Richards MA. Cancer  
52  
53 384 recording and mortality in the General Practice Research Database and linked cancer  
54  
55 385 registries. *Pharmacoepidemiol Drug Saf* 2013. DOI:10.1002/pds.3374.  
56  
57  
58 386 7 Rañopa M, Douglas I, van Staa T, *et al*. The identification of incident cancers in UK primary  
59  
60

- 1  
2  
3 387 care databases: a systematic review. *Pharmacoepidemiol Drug Saf* 2015; **24**: 11–8.  
4  
5  
6 388 8 Arhi CS, Bottle A, Burns EM, *et al*. Comparison of cancer diagnosis recording between the  
7  
8 389 Clinical Practice Research Datalink, Cancer Registry and Hospital Episodes Statistics. *Cancer*  
9  
10 390 *Epidemiol* 2018; **57**: 148–57.  
11  
12  
13 391 9 UK and Ireland Association of Cancer Registries. UK and Ireland Association of Cancer  
14  
15 392 Registries. <http://www.ukiacr.org/kpis/> (accessed March 18, 2019).  
16  
17  
18 393 10 Di Girolamo C, Walters S, Benitez Majano S, *et al*. Characteristics of patients with missing  
19  
20 394 information on stage: a population-based study of patients diagnosed with colon, lung or  
21  
22 395 breast cancer in England in 2013. *BMC Cancer* 2018; **18**: 492.  
23  
24  
25  
26 396 11 Barclay ME, Lyratzopoulos G, Greenberg DC, Abel GA. Missing data and chance variation in  
27  
28 397 public reporting of cancer stage at diagnosis: Cross-sectional analysis of population-based  
29  
30 398 data in England. *Cancer Epidemiol* 2018; **52**: 28–42.  
31  
32  
33 399 12 Herbert A, Wijlaars L, Zylbersztejn A, Cromwell D, Hardelid P. Data Resource Profile: Hospital  
34  
35 400 Episode Statistics Admitted Patient Care (HES APC). *Int J Epidemiol* 2017; **46**: 1093-1093i.  
36  
37  
38 401 13 Public Health England. Cancer Outcome and Services Data set - User guide v9.0.3. 2019.  
39  
40 402 [http://www.ncin.org.uk/collecting\\_and\\_using\\_data/data\\_collection/cosd](http://www.ncin.org.uk/collecting_and_using_data/data_collection/cosd) (accessed March  
41  
42 403 26, 2020).  
43  
44  
45  
46 404 14 Shah A, Rachet B, Mistry E, Cooper N, Brown CM, Coleman MP. Survival from bladder cancer in  
47  
48 405 england and wales up to 2001. *Br J Cancer* 2008; **99**: S86–9.  
49  
50  
51 406 15 Lewis JD, Bilker WB, Weinstein RB, Strom BL. The relationship between time since registration  
52  
53 407 and measured incidence rates in the General Practice Research Database.  
54  
55 408 *Pharmacoepidemiol Drug Saf* 2005; **14**: 443–51.  
56  
57  
58 409 16 Møller H, Richards S, Hanchett N, *et al*. Completeness of case ascertainment and survival time  
59  
60

- 1  
2  
3 410 error in English cancer registries: Impact on 1-year survival estimates. *Br J Cancer* 2011; **105**:  
4  
5 411 170–6.  
6  
7  
8 412 17 Wolf A, Dedman D, Campbell J, *et al.* Data resource profile: Clinical Practice Research Datalink  
9  
10 413 (CPRD) Aurum. *Int J Epidemiol* 2019; **48**: 1740-1740g.  
11  
12  
13 414 18 Bright CJ, Lawton S, Benson S, *et al.* Data Resource Profile: The Systemic Anti-Cancer Therapy  
14  
15 415 (SACT) Dataset. *Int J Epidemiol* 2020; **49**: 15-15l.  
16  
17  
18 416  
19  
20  
21 417  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60

For peer review only



**Table 1:** Time in days from main gold standard diagnosis date to first ever record in each combination of sources

Cancer	NCRAS		NCRAS & HES APC		CPRD GOLD, HES APC & ONS MORTALITY		CPRD GOLD		HES APC	
	median (IQR)	5th-95th percentile	median (IQR)	5th-95th percentile	median (IQR)	5th-95th percentile	median (IQR)	5th-95th percentile	median (IQR)	5th-95th percentile
Oral Cavity (C00-06)	0 (0, 0)	0-20	0 (0, 0)	0-12	0 (0, 17)	0-57	11 (0, 25)	0-80	12 (0, 39)	0-91
Oesophageal (C15)	0 (0, 1)	0-30	0 (0, 0)	0-6	0 (0, 0)	0-30	7 (0, 18)	0-59	0 (0, 6)	0-85
Stomach (C16)	0 (0, 2)	0-28	0 (0, 0)	0-0	0 (0, 0)	0-37	10 (1, 22)	0-64	0 (0, 0)	0-64
Colorectal (C18-C20)*	0 (0, 3)	0-41	0 (0, 0)	0-19	0 (0, 0)	0-36	7 (0, 21)	0-70	0 (0, 15)	0-90
Liver (C22)	0 (0, 7)	0-87	0 (0, 0)	0-51	0 (0, 2)	0-72	9 (0, 29)	0-113	0 (0, 32)	0-170
Pancreas (C25)	0 (0, 8)	0-56	0 (0, 0)	0-23	0 (0, 0)	0-52	8 (0, 22)	0-76	0 (0, 8)	0-101
Lung (C34)*	0 (0, 5)	0-42	0 (0, 0)	0-20	0 (0, 4)	0-56	10 (0, 22)	0-85	0 (0, 19)	0-190
Malignant melanoma (C43)	0 (0, 0)	0-23	0 (0, 0)	0-29	0 (0, 21)	0-64	11 (0, 25)	0-73	31 (0, 61)	0-240
Breast (C50)*	0 (0, 0)	0-26	0 (0, 0)	0-27	7 (0, 14)	0-37	7 (0, 14)	0-48	27 (16, 41)	0-365
Cervix (C53)	0 (0, 0)	0-17	0 (0, 0)	0-3	3 (0, 20)	0-74	13 (4, 27)	0-79	17 (0, 48)	0-113
Uterus (C54-55)	0 (0, 0)	0-19	0 (0, 0)	0-4	0 (0, 19)	0-55	14 (7, 27)	0-69	8 (0, 41)	0-89
Ovaries (C56)	0 (0, 3)	0-33	0 (0, 0)	0-21	0 (0, 0)	0-41	10 (0, 24)	0-95	0 (0, 14)	0-96
Prostate (C61)*	0 (0, 0)	0-68	0 (0, 0)	0-82	2 (0, 22)	0-154	15 (3, 29)	0-112	65 (0, 423)	0-2,113
Kidney (C64)	0 (0, 5)	0-66	0 (0, 0)	0-36	0 (0, 0)	-24-78	0 (0, 22)	0-112	0 (0, 20)	0-250
Bladder (C67)	1 (0, 15)	0-222	0 (0, 0)	0-31	0 (0, 0)	0-29	7 (0, 30)	0-166	0 (0, 0)	0-99
Brain/CNS (C71-72)	1 (0, 8)	0-63	0 (0, 0)	0-31	0 (0, 0)	0-32	8 (0, 20)	0-68	0 (0, 1)	0-166
Thyroid (C73)	0 (0, 0)	0-28	0 (0, 0)	0-20	0 (0, 25)	0-87	22 (3, 42)	0-127	1 (0, 58)	0-154
Non-Hodgkin lymphoma (C82-85)	0 (0, 3)	0-43	0 (0, 0)	0-33	0 (0, 12)	0-61	16 (4, 32)	0-118	0 (0, 31)	0-551
Multiple myeloma (C90)	0 (0, 8)	0-235	0 (0, 0)	0-80	0 (0, 1)	0-75	10 (0, 28)	0-148	0 (0, 41)	0-714
Leukemia (C91-95)	0 (0, 7)	0-909	0 (0, 1)	0-1,038	0 (0, 0)	0-89	1 (0, 20)	0-140	0 (0, 180)	0-1,811

Footnote: Number of days between main gold standard diagnosis date and applied definitions. Cancer sites are ordered according to corresponding codes from the International Classification of Diseases, version 10

(ICD-10). \*Four most common cancer sites. All sources definition not shown as diagnosis date is the same as the gold standard definition by default. NCRAS = National Cancer Registration and Analysis Service cancer

registration data. CPRD = Clinical Practice Research Datalink. HES APC = Hospital Episode Statistics Admitted Patient Care data. ONS = Office for National Statistics

1  
2  
3 **Figure 1:**  
4

5  
6 Title: Gold standard algorithm to identify incident site-specific cancers using all data sources  
7

8  
9 **Figure 2:**  
10

11 Title: Positive Predictive Value of cancer diagnoses for each combination of sources when compared  
12  
13  
14 to the main gold standard algorithm  
15

16  
17 Legend: Percentage of incident cancers defined using the first ever record in each combination of  
18  
19 sources confirmed by a gold standard algorithm that considers confirmatory and contradictory data  
20  
21 from each source. Cancer sites are ordered according to corresponding codes from the International  
22  
23 Classification of Diseases, version 10 (ICD-10). \*Four most common cancer sites. NCRAS = National  
24  
25 Cancer Registration and Analysis Service cancer registration data. CPRD = Clinical Practice Research  
26  
27 Datalink. HES APC = Hospital Episode Statistics Admitted Patient Care data. ONS = Office for National  
28  
29 Statistics  
30  
31  
32

33 **Figure 3:**  
34

35  
36 Title: Sensitivity of cancer diagnoses for each combination of sources when compared to the main  
37  
38 gold standard algorithm  
39

40  
41 Legend: Percentage of incident cancers identified using the main gold standard algorithm that  
42  
43 considers confirmatory and contradictory data from each source that are identified using the first  
44  
45 ever record in each combination of sources. Cancer sites are ordered according to corresponding  
46  
47 codes from the International Classification of Diseases, version 10 (ICD-10). \*Four most common  
48  
49 cancer sites. NCRAS = National Cancer Registration and Analysis Service cancer registration data.  
50  
51 CPRD = Clinical Practice Research Datalink. HES APC = Hospital Episode Statistics Admitted Patient  
52  
53 Care data. ONS = Office for National Statistics  
54  
55  
56

57  
58  
59 **Figure 4:**  
60

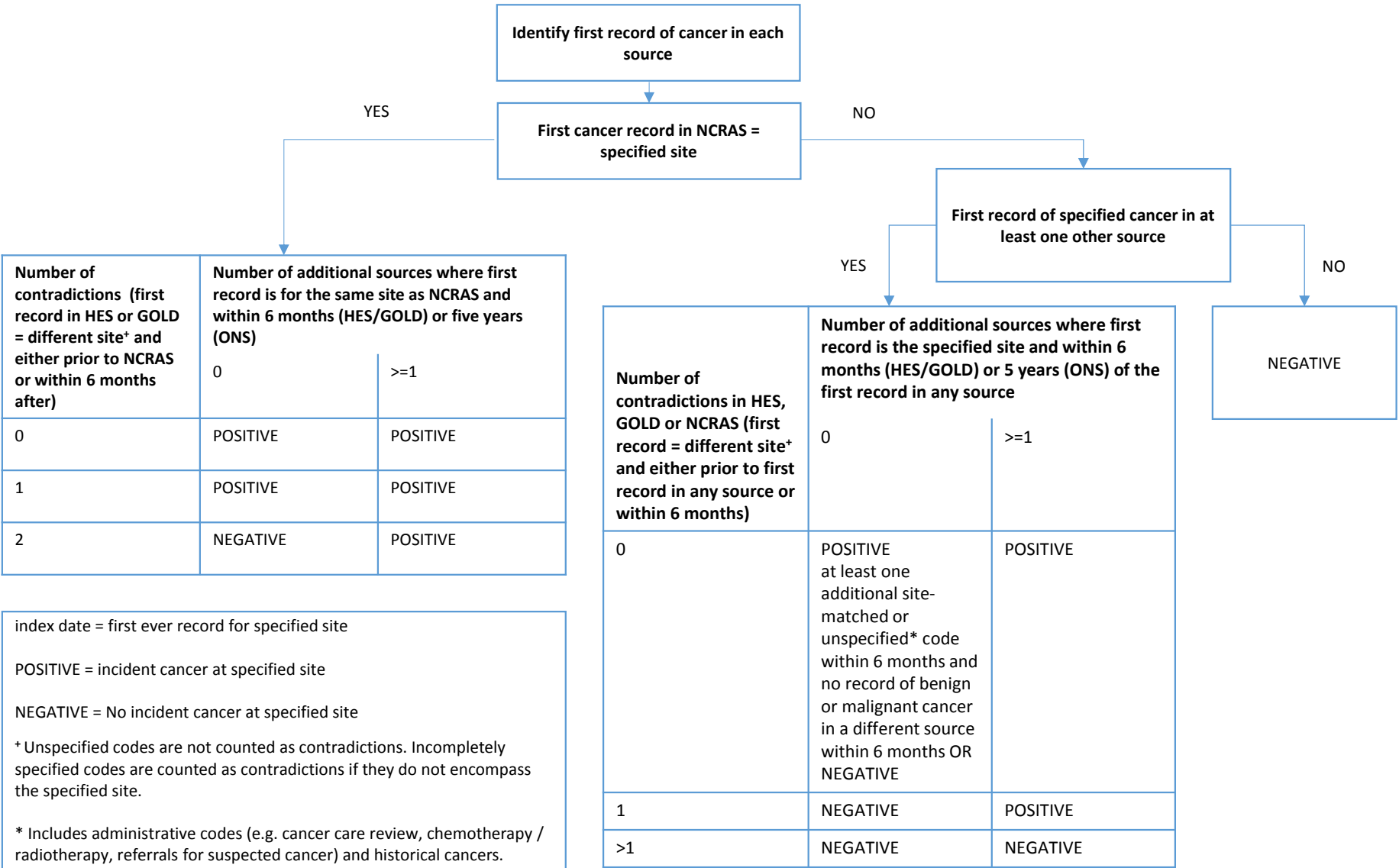
1  
2  
3 Title: Mortality following first ever record of cancer in each combination of sources  
4  
5

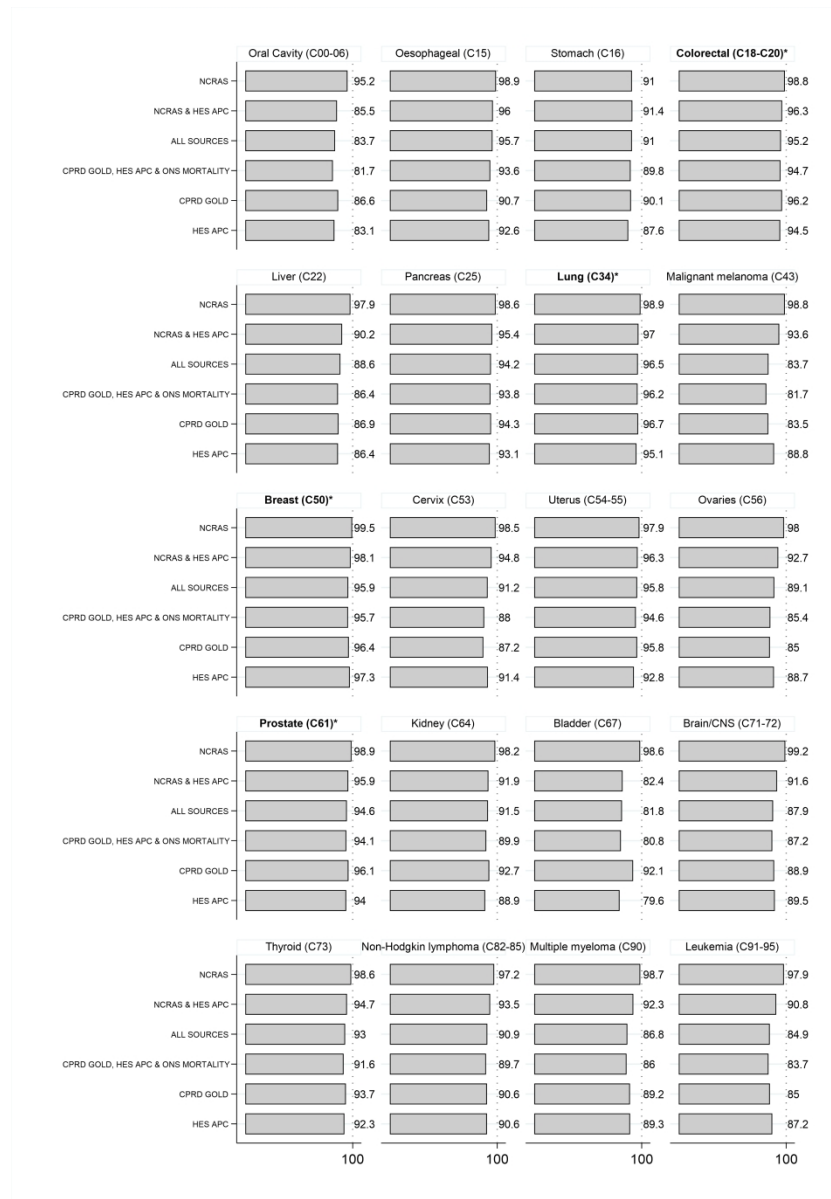
6 Legend: Cancer sites are ordered according to corresponding codes from the International  
7  
8 Classification of Diseases, version 10 (ICD-10). \*Four most common cancer sites. NHL = Non hodgkin  
9  
10 lymphoma. NCRAS = National Cancer Registration and Analysis Service cancer registration data.  
11  
12 CPRD = Clinical Practice Research Datalink. HES APC = Hospital Episode Statistics Admitted Patient  
13  
14 Care data. ONS = Office for National Statistics  
15  
16

17  
18 **Figure 5:**

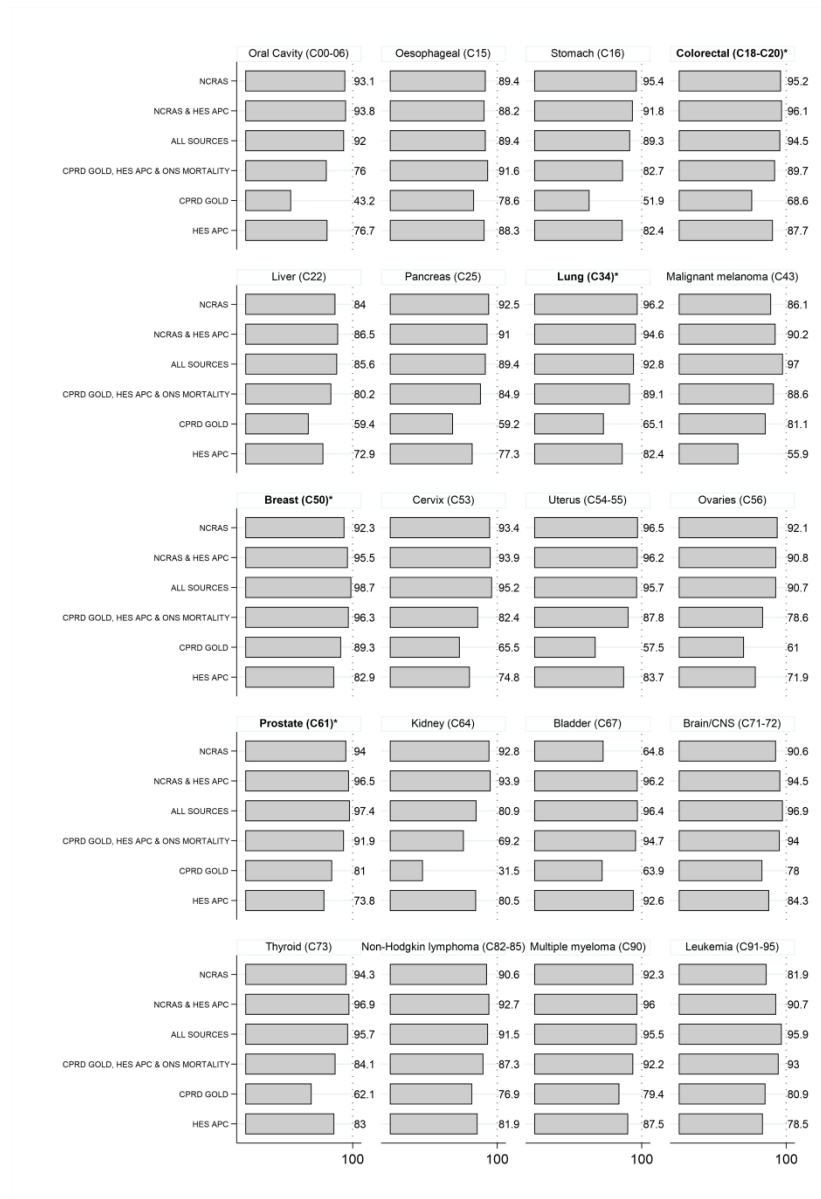
19  
20  
21 Title: Completeness of grade and stage for cancers identified using NCRAS data only  
22

23  
24 Legend: Cancer sites are ordered according to corresponding codes from the International  
25  
26 Classification of Diseases, version 10 (ICD-10). \*Four most common cancer sites. NHL = Non hodgkin  
27  
28 lymphoma. Grading information is not applicable to brain/CNS, sarcoma or haematological cancers  
29  
30 and not required by in the national data standard (COSD) for prostate cancer. Core staging is not  
31  
32 applicable to haematological and gynaecological cancers. Other types of staging are recommended  
33  
34 by COSD.  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60



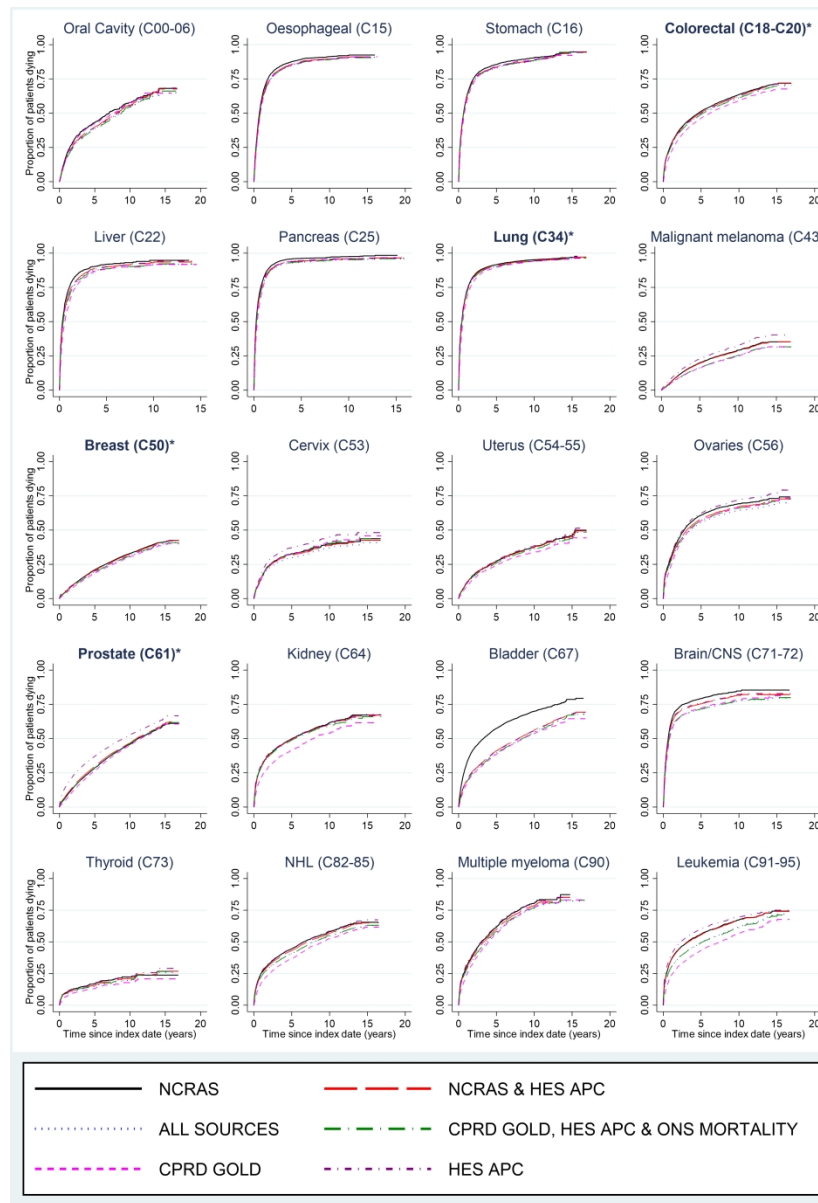


Title: Figure 2: Positive Predictive Value of cancer diagnoses for each combination of sources when compared to the main gold standard algorithm. Legend: Percentage of incident cancers defined using the first ever record in each combination of sources confirmed by a gold standard algorithm that considers confirmatory and contradictory data from each source. Cancer sites are ordered according to corresponding codes from the International Classification of Diseases, version 10 (ICD-10). \*Four most common cancer sites. NCRAS = National Cancer Registration and Analysis Service cancer registration data. CPRD = Clinical Practice Research Datalink. HES APC = Hospital Episode Statistics Admitted Patient Care data. ONS = Office for National Statistics

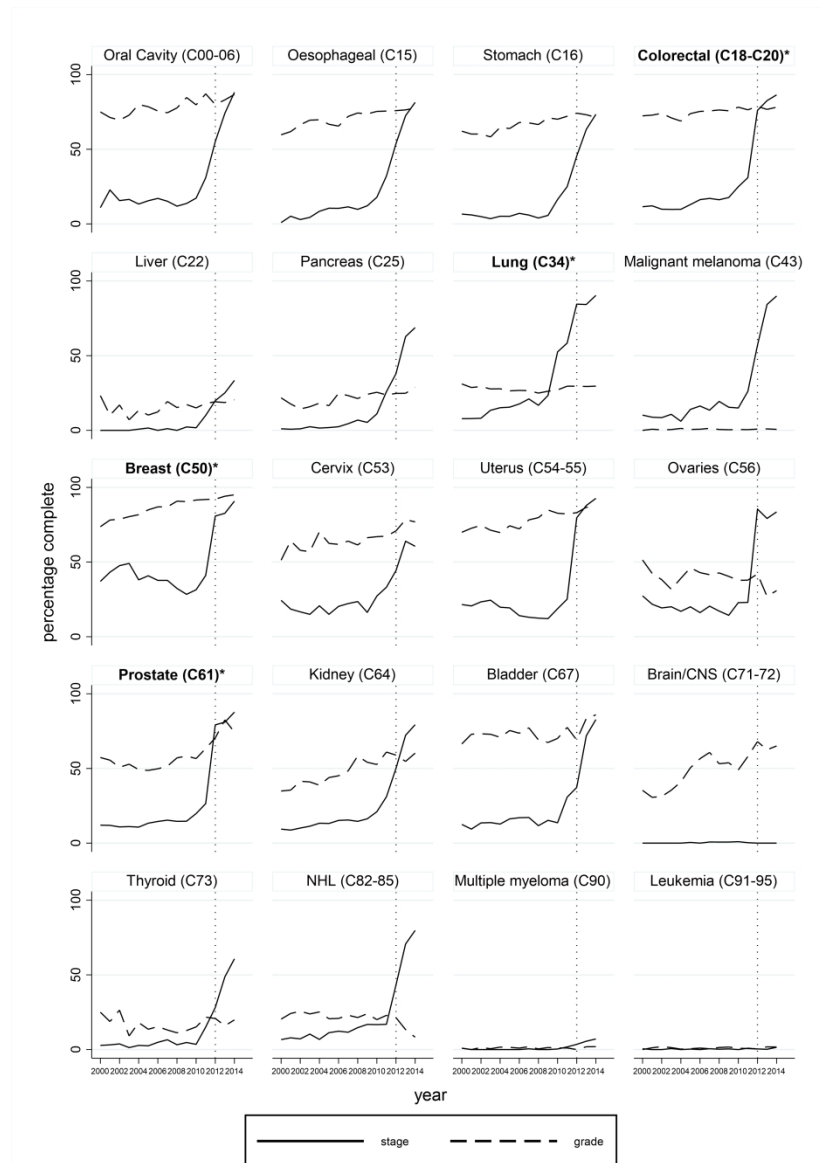


Title: Figure 3: Sensitivity of cancer diagnoses for each combination of sources when compared to the main gold standard algorithm  
 Legend: Percentage of incident cancers identified using the main gold standard algorithm that considers confirmatory and contradictory data from each source that are identified using the first ever record in each combination of sources. Cancer sites are ordered according to corresponding codes from the International Classification of Diseases, version 10 (ICD-10). \*Four most common cancer sites.

NCRAS = National Cancer Registration and Analysis Service cancer registration data. CPRD = Clinical Practice Research Datalink. HES APC = Hospital Episode Statistics Admitted Patient Care data. ONS = Office for National Statistics



Title: Figure 4: Mortality following first ever record of cancer in each combination of sources  
 Legend: Cancer sites are ordered according to corresponding codes from the International Classification of Diseases, version 10 (ICD-10). \*Four most common cancer sites. NHL = Non hodgkin lymphoma. NCRAS = National Cancer Registration and Analysis Service cancer registration data. CPRD = Clinical Practice Research Datalink. HES APC = Hospital Episode Statistics Admitted Patient Care data. ONS = Office for National Statistics



Title: Figure 5: Completeness of grade and stage for cancers identified using NCRAS data only  
 Legend: Cancer sites are ordered according to corresponding codes from the International Classification of Diseases, version 10 (ICD-10). \*Four most common cancer sites. NHL = Non hodgkin lymphoma. Grading information is not applicable to brain/CNS, sarcoma or haematological cancers and not required by in the national data standard (COSD) for prostate cancer. Core staging is not applicable to haematological and gynaecological cancers. Other types of staging are recommended by COSD.



## Supplementary appendix

### Benefits and limitations of using individual and different combinations of linked English routine data sources in cancer epidemiology studies

Table 1: Number of patients identified with each cancer site using the gold standard algorithm

Cancer site	Number of patients
Oral Cavity (C00-06)	2097
Oesophageal (C15)	5212
Stomach (C16)	4016
Colorectal (C18-C20)*	22173
Liver (C22)	2230
Pancreas (C25)	5008
Lung (C34)	21978
Malignant melanoma (C43)	7282
Breast (C50)	29297
Cervix (C53)	1503
Uterus (C54-55)	4325
Ovaries (C56)	4157
Prostate (C61)	24888
Kidney (C64)	4086
Bladder (C67)	8871
Brain/CNS (C71-72)	2921
Thyroid (C73)	1314
NHL (C82-85)	6644
Multiple myeloma (C90)	2672
Leukemia (C91-95)	5279
Total	165953

Figure 1: Positive Predictive Value by age

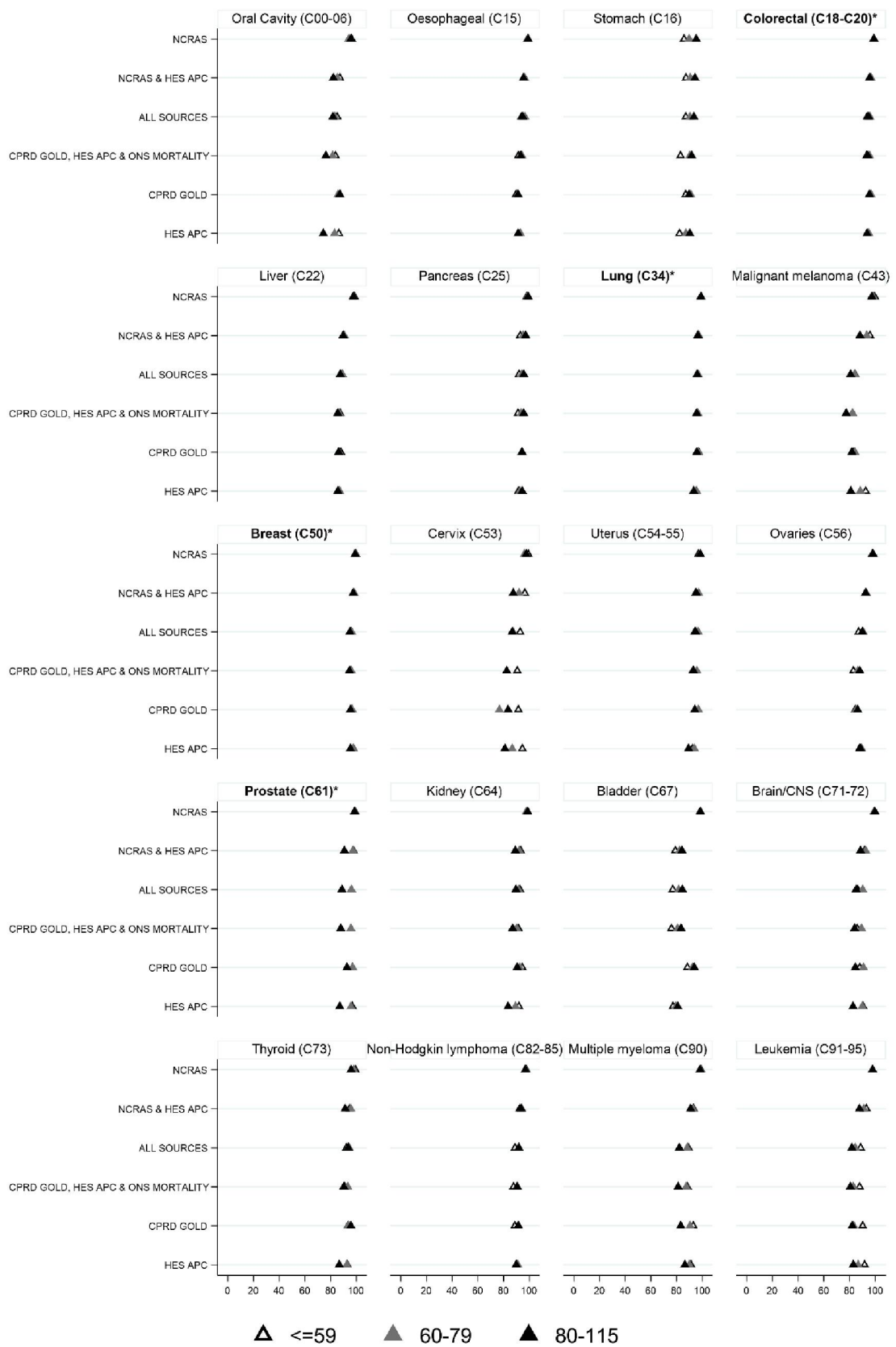


Figure 2: Positive Predictive Value by sex

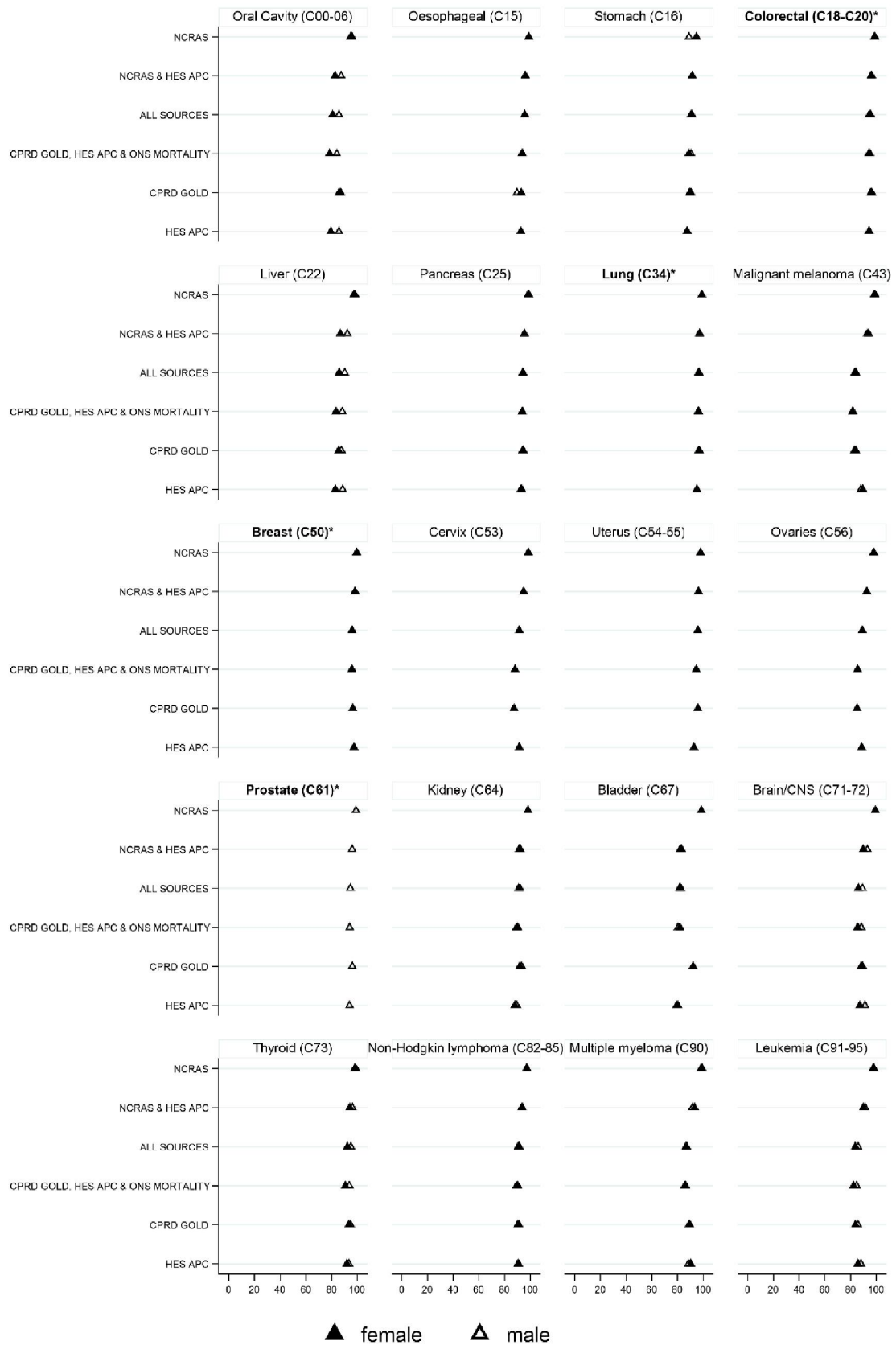


Figure 3: Positive Predictive Value by calendar year

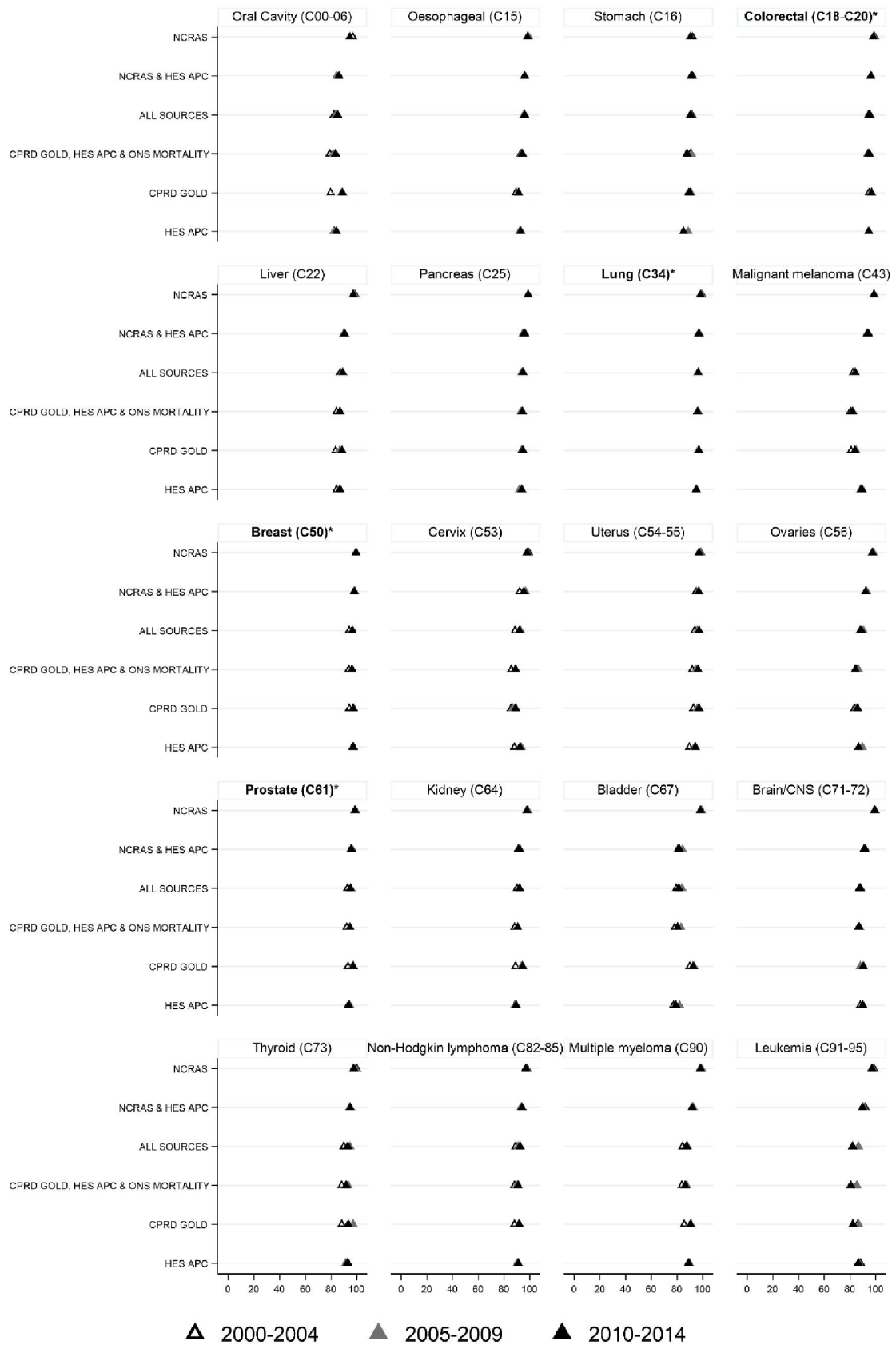


Figure 4: Sensitivity by age

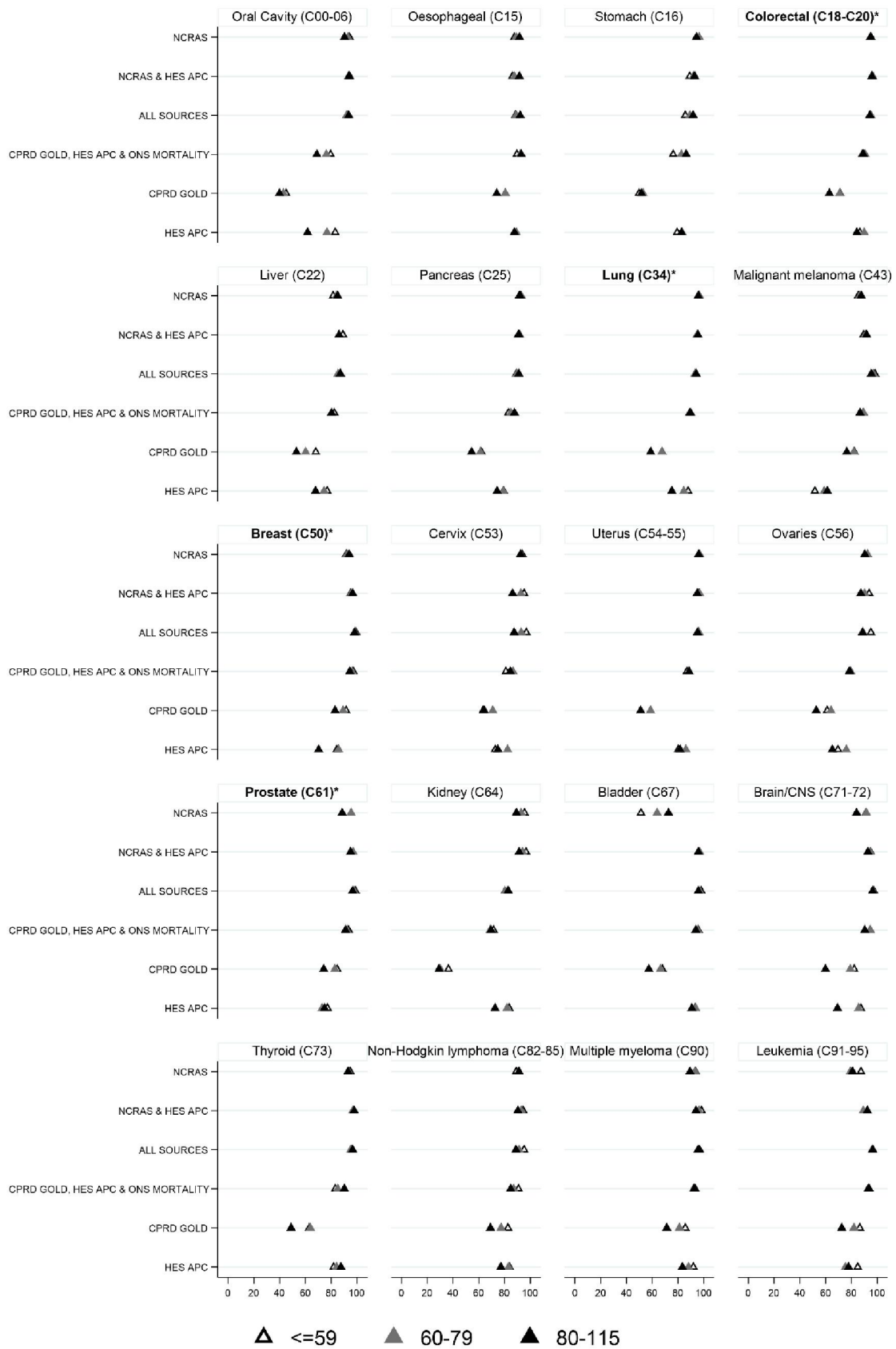


Figure 5: Sensitivity by sex

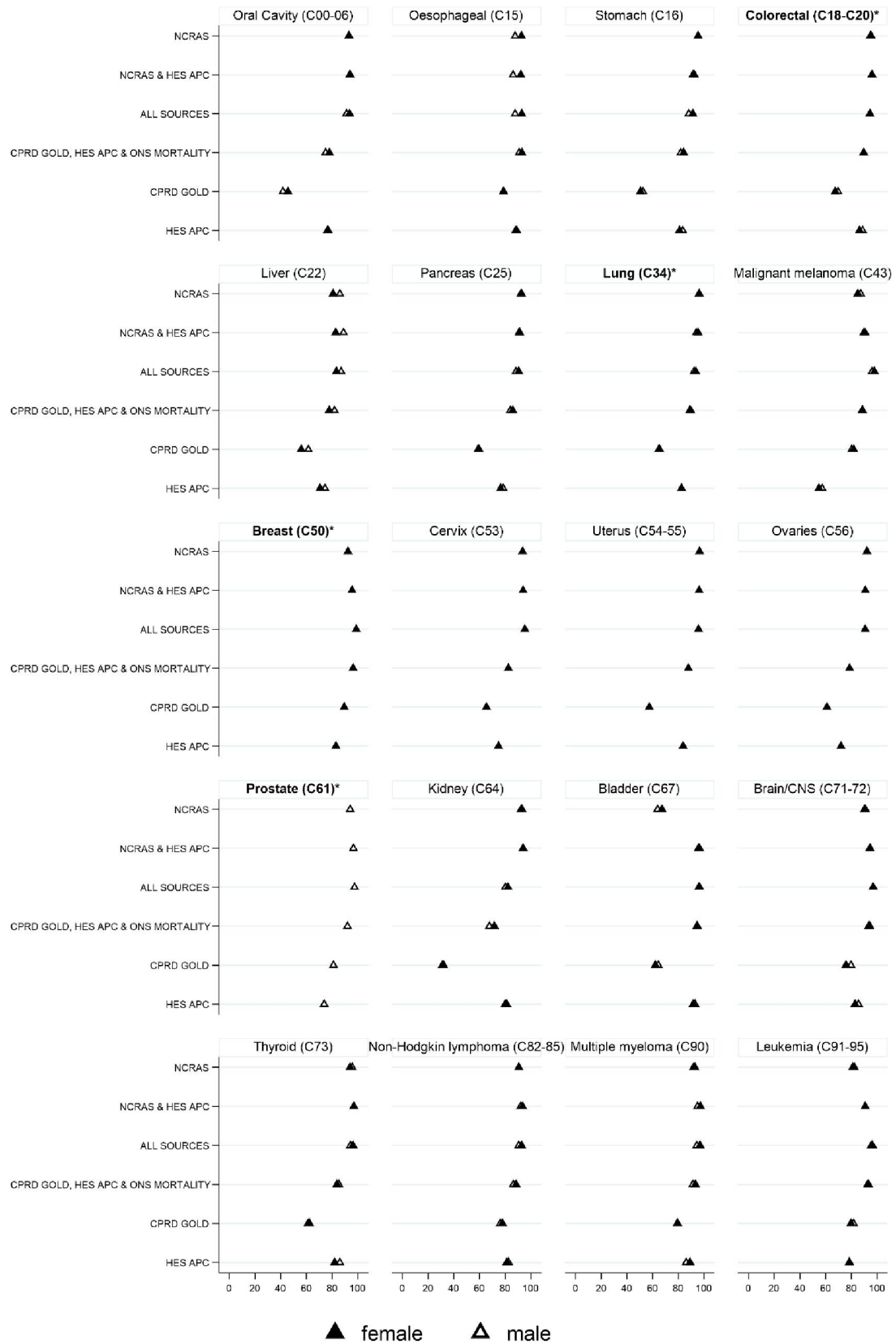


Figure 6: Sensitivity by calendar year

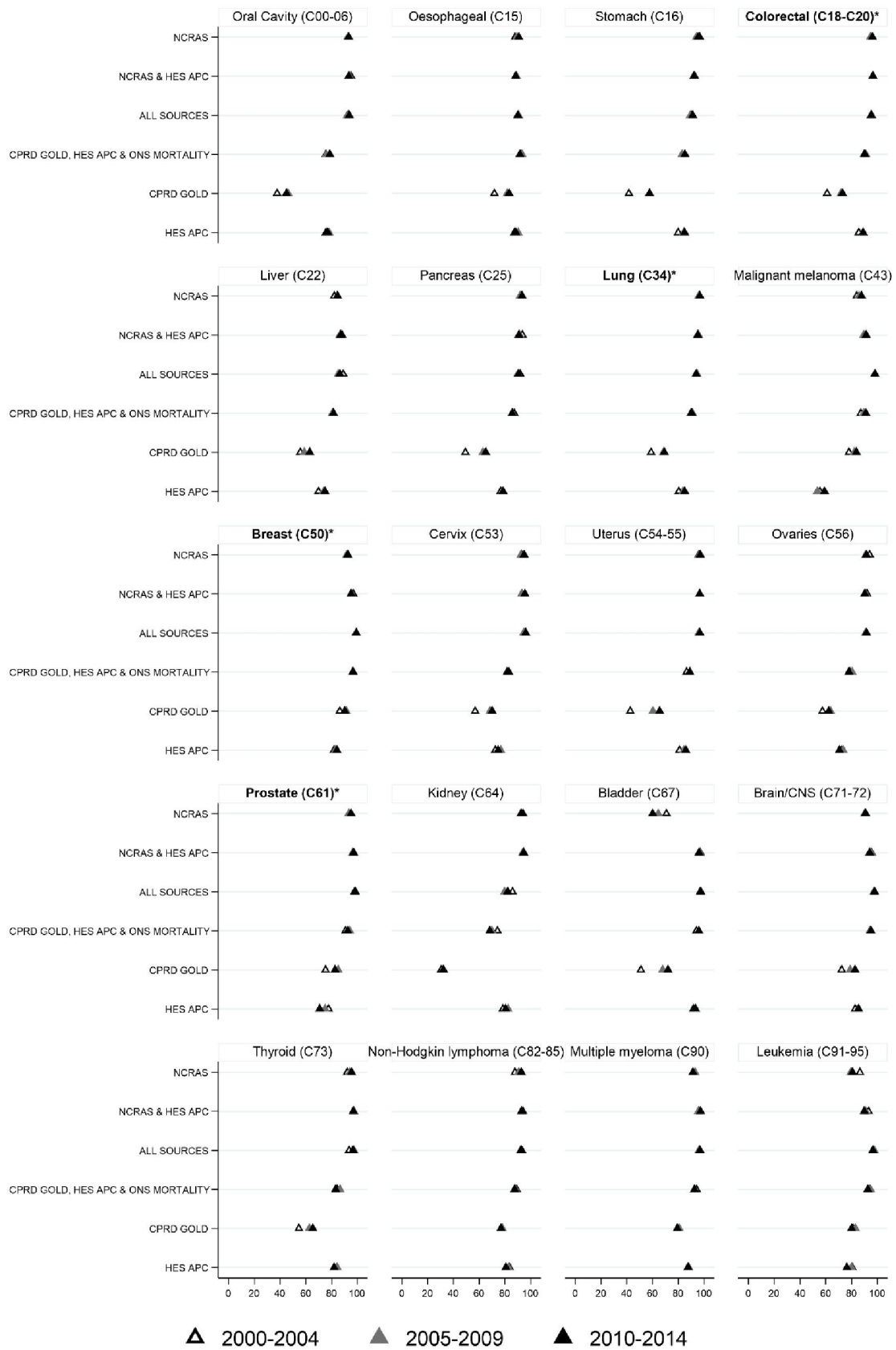


Figure 7: Output from logistic regression models with completeness of stage and grade as the dependent variables

Created using coefplot command in Stata <http://repec.sowi.unibe.ch/stata/coefplot/>

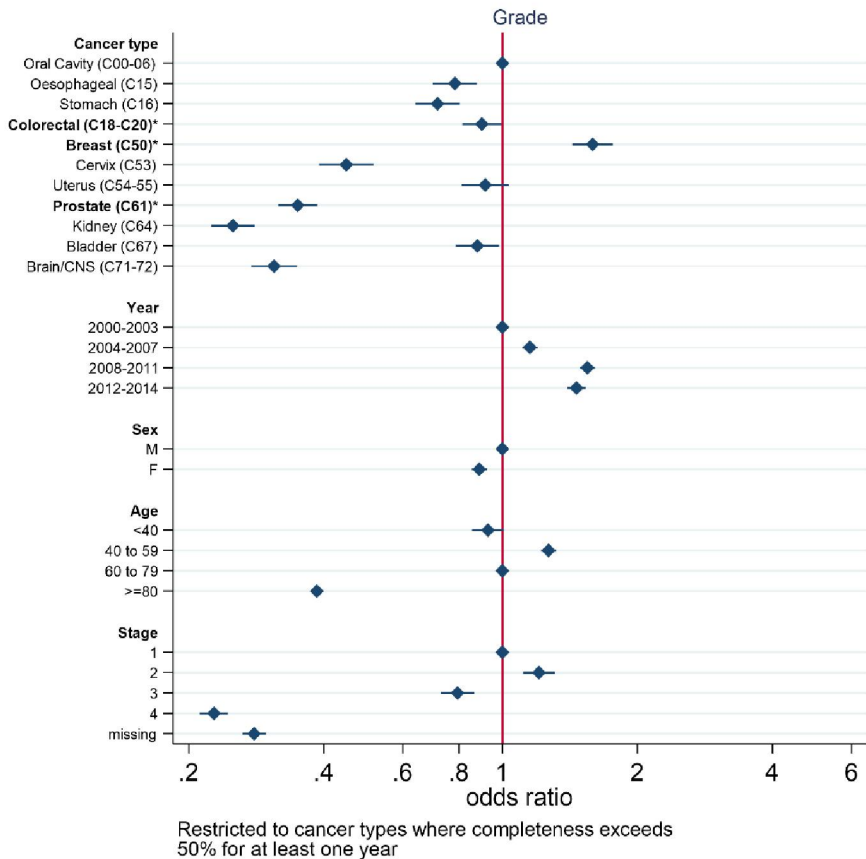
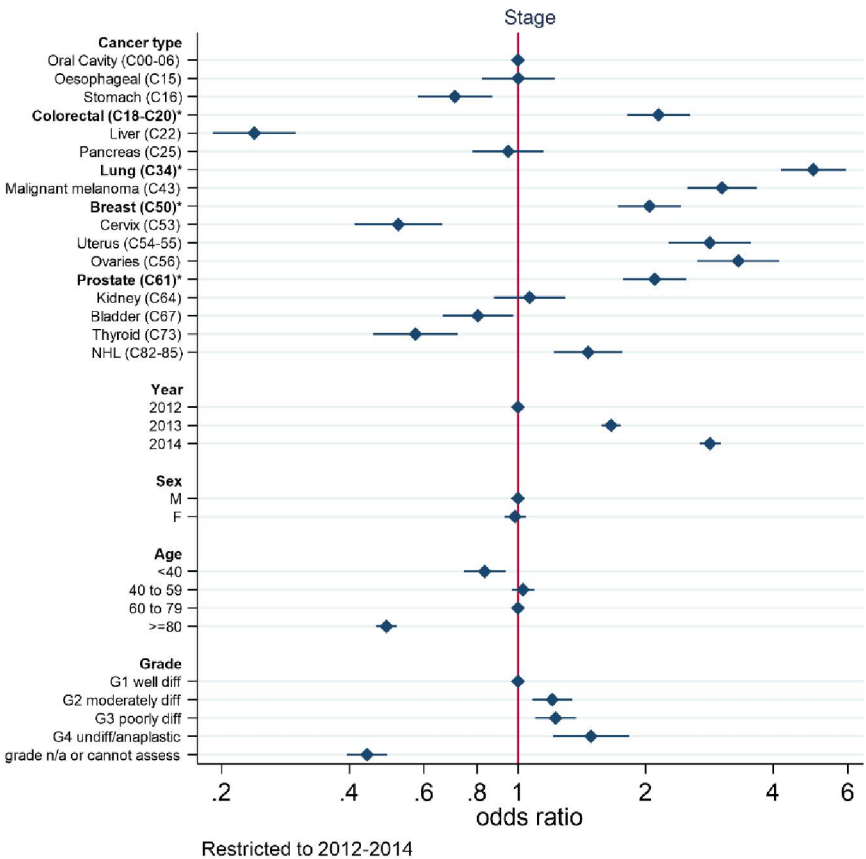




Figure 8: Recording of treatment modalities for patients identified using NCRAS data only

