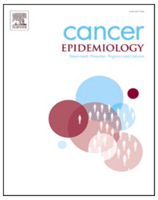
Cancer Epidemiology 94 (2025) 102703

Contents lists available at ScienceDirect

Cancer Epidemiology

journal homepage: www.elsevier.com/locate/canep



Differences in recording of cancer diagnosis between datasets in England: A population-based study of linked cancer registration, hospital, and primary care data

Emma Whitfield [a,b,*,1], Becky White [a,2], Matthew E. Barclay [a,3], Meena Rafiq [a,4], Cristina Renzi [a,c,5], Brian Rous [d,6], Spiros Denaxas [b,e,f,g,7], Georgios Lyratzopoulos [a,8]

[a] *ECHO (Epidemiology of Cancer Healthcare & Outcomes), Department of Behavioural Science and Health, Institute of Epidemiology and Health Care, UCL (University College London), 1-19 Torrington Place, London WC1E 7HB, UK*
[b] *Institute of Health Informatics, UCL, London, UK*
[c] *Faculty of Medicine, University Vita-Salute San Raffaele, Milan, Italy*
[d] *National Cancer Registration and Analysis Service, NHS England, London, UK*
[e] *British Heart Foundation Data Science Centre, London, UK*
[f] *Health Data Research UK, London, UK*
[g] *UCL Hospitals Biomedical Research Centre, London, UK*

ARTICLE INFO

Keywords:
Registry
Medical records

ABSTRACT

Background: Differences in the recording of cancer case status and diagnosis date have been observed between cancer registry (CR) – the reference standard – and electronic health records (EHRs); such differences may affect estimates of cancer risk or misclassify diagnostic pathways. This study aims to quantify differences in recording of case status and date of cancer diagnosis between cancer registry and EHRs.

Methods: Linked primary care (Clinical Practice Research Datalink (CPRD)), secondary care (Hospital Episode Statistics (HES)) and national Cancer Registry (CR) data, were used to identify 14,301 patients with a recorded diagnosis of brain, colon, lung, ovarian, or pancreatic cancer between 1999 and 2018. Agreement in case status between datasets, differences in recorded diagnosis dates, and change in agreement over time were investigated for each cancer site.

Results: Between 84 % (ovary) to 92 % (colon) of diagnoses in cancer registry were also recorded in combined CPRD-HES data. Agreement with cancer registry was slightly lower in HES (78 % (ovary) to 86 % (colon)) and CPRD (61 % (ovary, pancreas) to 72 % (brain)). The proportion of CPRD-HES diagnoses confirmed in CR varied by cancer site (50 % (brain) to 86 % (lung)). Agreement between CR and HES was relatively stable within cancer sites over time. Concordance between CR and CPRD was more heterogeneous between cancer sites and over time. Best agreement in diagnosis date was observed between CR and HES (median difference 0 or 1 days, all cancer sites).

Conclusion: Agreement between CR and EHR data is heterogeneous across cancer sites. Concordance does not appear to have improved over time. Combined data from primary and secondary care may be sufficient to

* Corresponding author at: ECHO (Epidemiology of Cancer Healthcare & Outcomes), Department of Behavioural Science and Health, Institute of Epidemiology and Health Care, UCL (University College London), 1-19 Torrington Place, London WC1E 7HB, UK.

E-mail addresses: emma.whitfield.20@ucl.ac.uk (E. Whitfield), becky.white.19@ucl.ac.uk (B. White), m.barclay@ucl.ac.uk (M.E. Barclay), meena.rafiq@ucl.ac.uk (M. Rafiq), c.renzi@ucl.ac.uk (C. Renzi), brian.rous@nhs.net (B. Rous), s.denaxas@ucl.ac.uk (S. Denaxas), y.lyratzopoulos@ucl.ac.uk (G. Lyratzopoulos).

[1] ORCiD: 0000–0001-5427–7150
[2] ORCiD: 0000–0002-0643–7890
[3] ORCiD: 0000–0003-1148–1922
[4] ORCiD: 0000–0002-1837–1542
[5] ORCiD: 0000–0003-3845–9493
[6] ORCiD: 0000–0002-7619–461X
[7] ORCiD: 0000–0001-9612–7791
[8] ORCiD: 0000–0002-2873–7421

https://doi.org/10.1016/j.canep.2024.102703
Received 13 June 2024; Received in revised form 6 November 2024; Accepted 10 November 2024
Available online 29 November 2024

approximate case status in CR in some circumstances, but the date we consider to represent the diagnosis may impact study outcomes.

1. Introduction

Many studies need to reliably identify cancer cases and their diagnosis dates. Population-based cancer registries (CR) are considered the 'reference standard' for this purpose, however routinely-collected hospital data and electronic health records (EHR) are also used, when cancer registry data do not exist, are unavailable for linkage, or access is limited or delayed (Supplementary Information, Supp. Table 1 [1,2]). The dates recorded in these data sources are derived for different purposes but have all been used in research as a proxy for the cancer diagnosis date [3–5] – we refer to these as the 'cancer date' from here.

In the UK, discrepancies in the recording of cancer diagnosis – mismatching case status and/or cancer date – have been observed between cancer registry data and primary and secondary care EHRs [6–10]. It is unclear how these differences vary over time and across cancer sites, and how this may impact the results of studies which rely on accurately determining cancer diagnoses. For instance, estimates of the positive predictive value of presenting cancer symptoms may be biased due to artefactual differences in case status.

This study aims to describe differences in the recording of cancer diagnoses in primary and secondary care EHRs, and a population-based cancer registration system in England. A secondary aim is to determine whether differences in case status between data sources have changed over time.

2. Methods

2.1. Data sources

This study used data from the Clinical Practice Research Datalink (CPRD) GOLD, linked to National Cancer Registration and Analysis Service cancer registration (CR) and Hospital Episode Statistics (HES) Admitted Patient Care (APC) datasets.

CPRD GOLD is a database of anonymized routinely-collected primary care data from participating UK GP practices using Vision software [11]. HES APC captures data on all admissions to English NHS providers from April 1997 – March 2021 [12]. The population-based CR dataset contains data on all tumours diagnosed in England between January 1990 – December 2018 [13,14]. To achieve this, patient-level data is collated from multiple sources such as hospital records (including HES), treatment records, pathology reports, and multidisciplinary team meetings [14]. Cancer registration officers process this data, seeking additional information from primary or secondary care if necessary, and then follow international standards to determine the 'date of incidence' of each tumour [13,15]. Data from HES and CR were linked to CPRD GOLD using an eight-step deterministic linkage algorithm based on NHS number, sex, date of birth, and postcode [12,16].

2.2. Study population

This study investigated agreement in recorded cancer diagnosis for five exemplar cancer sites (brain - including benign tumours, colon, lung, ovarian, and pancreatic cancer). Codelists of Read v2 (for CPRD) and ICD-10 codes (for CR and HES) were developed for each cancer site (see Supplementary Information) [17].

We selected cases from a random sample of 1 million CPRD GOLD patients who were registered at an up-to-standard CPRD practice for at least one year between 1/1/2007–31/10/2021, whilst aged 30–99, and who were eligible for linkage to HES and CR data. From this sample we included any patient who had a recorded diagnosis of brain, colon, lung, ovarian, or pancreatic cancer in any of CPRD, HES, or CR.

For each cancer site, for each of CR, CPRD, and HES APC, we identified patients with a diagnosis code (Read v2/ ICD-10) for the selected site recorded between 1/1/1999 and 31/12/2018. For this study, for each patient we only considered data from all three datasets when patients were fully eligible for study in CPRD. Detailed inclusion criteria are described in the Supplementary Information.

To compare the use of combined primary and secondary care records to cancer registry data, we applied the same selection criteria to combined records from CPRD GOLD and HES APC to identify a fourth set of cases– referred to as the EHR case set.

Cancer dates were defined as the date of the recorded cancer diagnosis in CPRD, the European Network of Cancer Registries (ENCR) guideline date of incidence in CR [15], and the start date of the consultant episode in which the diagnosis was recorded in HES APC.

2.3. Statistical analysis

2.3.1. Patient characteristics and agreement of cancer diagnosis

For each cancer site and dataset we report descriptive statistics for age at diagnosis, sex (both determined from CPRD), year of diagnosis, and Elixhauser Comorbidity Index at diagnosis (determined from HES APC) [18,19]; smoking status (determined from CPRD) is reported for lung cancer patients only (see Supplementary Information). Note that patient characteristics are reported to provide a descriptive summary of the patients captured in each dataset and statistical tests were not used to compare concordance of characteristics between datasets. Characteristics are determined from the datasets listed above regardless of the dataset in which the patient's diagnosis was recorded.

'Agreement of cancer diagnosis' between datasets was defined as identification of a case in both datasets, regardless of the timing of the cancer dates during the study period [6]. Agreement is reported between each pair of datasets for each cancer site.

To account for differences in tumour site coding between datasets, for each of CR, HES, and CPRD we further report the percentage of patients for each cancer site with any cancer diagnosis (i.e., including diagnoses at other cancer sites) recorded in each of the other datasets within one year.

2.3.2. Comparison of recorded cancer dates

We considered differences between cancer dates in the CR, as the 'reference standard', and other datasets. We restricted cases from the CR to only include diagnoses between 1/1/2000 and 31/12/2017, allowing identification of cases recorded up to one year earlier or later than the CR cancer date in each of the other datasets. We assumed that diagnoses recorded more than one year apart represented distinct diagnoses, as in Arhi et al. [6], and did not include them when calculating differences in cancer dates. For each cancer site and dataset we report the median and interquartile range of the difference in days between cancer dates, and the cumulative percentage of patients within each week of difference.

2.3.3. Changes in concordance with cancer registry over time

We used logistic regression to examine changes in concordance between the CR and CPRD or HES APC over time, stratifying by cancer site and adjusting for patient age and sex. We examined the probability of a diagnosis in CR being recorded in each of CPRD and HES APC both overall and (for CR patients diagnosed between 1/1/2000 and 31/12/2017) within one year. We modelled secular trends in agreement over diagnosis year using a cubic polynomial. Between 2013 and 2015 the eight English regional cancer registries were merged into a single national registry, establishing standardised data collection rules and a national data specification (COSD) for data reported from multidisciplinary teams [20]. To account for this we included dummy variables for pre-2013 (prior to change in cancer registration practices), 2014–15 (during change) and post-2016 (following change) to capture any step changes in our model.

2.3.4. Supplementary analysis

We investigated patient characteristics associated with the under-recording of diagnoses in CR. Cases diagnosed between 1/1/2014 and 31/12/2018 were identified from HES and CPRD, and for each dataset a logistic regression model was used to predict the likelihood of the diagnosis being recorded in CR, adjusting for cancer site, age at diagnosis, and Elixhauser Comorbidity Index at diagnosis.

3. Results

3.1. Patient characteristics

We identified 1731 brain, 4837 colon, 5420 lung 1406 ovarian, and 1280 pancreatic cancer patients with an eligible diagnosis recorded in at least one of CPRD, HES, or CR between 1/1/1999 and 31/12/2018 (Table 1). Case selection flow charts are given in Supplementary Information (Supp. Figs 1–5). For each cancer site, age, sex, year of diagnosis, and Elixhauser Comorbidity Index were broadly similar across data sources (Supplementary Information, Supp. Tables 2–6).

3.2. Agreement of cancer diagnosis

For all cancer sites, a diagnosis in CR was typically more likely to be recorded in HES than CPRD (Table 2). Agreement of cancer diagnosis ranged from 49.9 % (brain tumours, EHR diagnosis confirmed in CR) to 92.0 % (colon, CR diagnosis confirmed in EHR).The percentage of cases in combined EHR data recorded in CR data was much lower (ranging from 49.9 % (brain tumours) to 85.9 % (lung)) than the percentage of CR cases recorded in EHR data (Table 2).

Similar patterns were seen in the percentage of patients in each dataset with any cancer diagnosis recorded in another dataset within one year (Supp. Table 7). The percentage of CR-captured patients with any cancer diagnosis ranged from 77.6 % (brain tumours) to 87.3 % (colon) in CPRD, and from 86.8 % (ovarian) to 93.0 % (colon) in HES.

3.3. Comparison of recorded cancer dates

The proportion of cases with diagnoses recorded in CR (2000 – 2017) and a second dataset (1999 – 2018) with cancer dates that differed by more than one year ranged from 0.64 % (pancreatic, CR vs CPRD) to 5.34 % (brain, CR vs CPRD) (Supplementary Information, Supp. Table 8). For patients with diagnoses in CR and a second dataset within one year, differences in cancer date varied across cancer sites and datasets (Fig. 1, Fig. 2, Supplementary Information Supp. Tables 9–11).

3.4. Changes in concordance with cancer registry over time

The probability of a CR diagnosis being recorded in HES remained consistently high over time for lung cancer and pancreatic cancer (lung c.0.86, pancreatic – female patients c.0.87, pancreatic – male patients c.0.82), but was more variable for other cancer sites (Supplementary Information, Supp. Figures 6 – 10). The probability of a CR diagnosis being recorded in CPRD varied more over cancer sites and time, ranging from 0.46 (ovarian cancer 2000) to 0.77 (lung cancer, 2005–2006).

No statistically significant change was observed in the probability of recording for any cancer site in 2013, 2014–15, or 2016, although ovarian and brain cancer showed some volatility.

3.5. Supplementary analysis

In both HES and CPRD, patients with colon, lung, ovarian, and pancreatic cancers were significantly more likely to have their diagnosis recorded in CR than patients with brain tumours (Supplementary Information, Supp. Tables 12 & 13). HES patients with a higher Elixhauser Comorbidity Index were less likely to have their diagnosis recorded in CR.

4. Discussion

4.1. Summary

We examined differences in the recording of cancer diagnosis in England between three datasets widely used for research purposes. Of diagnoses recorded in the cancer registry, 84 % (ovarian) to 92 % (colon) were also recorded in combined EHRs, and recording was generally higher in secondary care. Of the cases recorded in primary care, 58 % (brain tumours) to 92 % (lung) were also recorded in the

Table 1
Summary statistics of patients with recorded cancer diagnoses in CR, CPRD, HES, and EHR (combined CPRD and HES). Patients are included once for each cancer site recorded in each dataset.

	CR (N=11,286)	CPRD (N=9415)	HES (N=11,750)	EHR (N=13,472)
Age				
Mean (SD)	71.1 (12.3)	70.1 (12.2)	70.6 (12.4)	70.6 (12.5)
Median (IQR)	72.0 (17.0)	71.0 (17.0)	72.0 (17.0)	72.0 (17.0)
Sex, n (%)				
Female	6000 (53.2 %)	5032 (53.5 %)	6253 (53.2 %)	7216 (53.6 %)
Male	5286 (46.8 %)	4383 (46.6 %)	5497 (46.8 %)	6256 (46.4 %)
Year of diagnosis, n (%)				
1999–2013	8583 (76.1 %)	7302 (77.6 %)	8961 (76.3 %)	10,323 (76.6 %)
2014–2018	2703 (24.0 %)	2113 (22.4 %)	2789 (23.7 %)	3149 (23.4 %)
Elixhauser score, n (%)				
0	5671 (50.3 %)	4604 (48.9 %)	6017 (51.2 %)	7088 (52.6 %)
1	2580 (22.9 %)	2306 (24.5 %)	2699 (23.0 %)	3035 (22.5 %)
2–3	2176 (19.3 %)	1822 (19.4 %)	2152 (18.3 %)	2384 (17.7 %)
4+	859 (7.61 %)	683 (7.25 %)	882 (7.51 %)	965 (7.16 %)
Cancer site, n (%)				
Brain	898 (7.96 %)	1122 (11.92 %)	1315 (11.2 %)	1606 (11.9 %)
Colon	3627 (32.1 %)	3210 (34.1 %)	3893 (33.1 %)	4517 (33.5 %)
Lung	4699 (41.6 %)	3522 (37.4 %)	4473 (38.1 %)	4952 (36.8 %)
Ovary	1027 (9.10 %)	811 (8.61 %)	1056 (8.99 %)	1231 (9.14 %)
Pancreas	1035 (9.17 %)	750 (7.97 %)	1013 (8.62 %)	1166 (8.65 %)

Table 2
Percentage of cases in original data source also recorded in other data source(s). EHR = combined CPRD and HES.

		Brain				Colon				Lung				Ovarian				Pancreatic			
		N = 1731				N = 4837				N = 5420				N = 1406				N = 1280			
		CR	CPRD	HES	EHR	CR	CPRD	HES	EHR	CR	CPRD	HES	EHR	CR	CPRD	HES	EHR	CR	CPRD	HES	EHR
Original data source	CR	-	72.0 (69.0, 74.9)	86.3 (83.8, 88.4)	89.2 (86.9, 91.1)	-	69.0 (67.4, 70.5)	86.4 (85.2, 87.5)	92.0 (91.1, 92.9)	-	68.5 (67.2, 69.9)	83.9 (82.8, 84.9)	90.5 (89.6, 91.3)	-	61.2 (58.2, 64.2)	77.5 (74.8, 80.0)	83.9 (81.5, 86.1)	-	61.2 (58.1, 64.1)	81.3 (78.7, 83.6)	89.2 (87.1, 91.0)
	CPRD	57.7 (54.7, 60.6)	-	70.5 (67.7, 73.1)	99.7 (99.2, 99.9)	77.9 (76.4, 79.3)	-	79.4 (78.0, 80.8)	99.5 (99.2, 99.7)	91.5 (90.5, 92.3)	-	85.7 (84.5, 86.8)	99.8 (99.5, 99.9)	77.6 (74.5, 80.4)	-	77.1 (74.0, 79.9)	99.4 (98.5, 99.8)	84.4 (81.6, 86.9)	-	79.2 (76.1, 82.0)	99.9 (99.1, 100.0)
	HES	58.9 (56.2, 61.6)	60.2 (57.4, 62.8)	-	97.2 (96.1, 98.0)	80.5 (79.2, 81.7)	65.5 (64.0, 67.0)	-	99.5 (9.2, 99.7)	88.1 (87.1, 89.0)	67.5 (66.1, 68.9)	-	99.6 (99.4, 99.8)	75.4 (82.6, 77.9)	59.2 (56.1, 62.2)	-	99.4 (98.7, 99.8)	83.0 (80.5, 85.3)	58.6 (55.5, 61.7)	-	99.8 (99.2, 100.0)
	EHR	49.9 (47.4, 52.3)	69.7 (67.4, 71.9)	79.6 (77.5, 81.5)	-	73.9 (72.6, 75.2)	70.7 (69.4, 72.0)	85.7 (84.7, 86.7)	-	85.9 (84.9, 86.8)	71.0 (69.7, 72.2)	90.0 (89.1, 90.8)	-	70.0 (67.4, 72.6)	65.5 (62.7, 68.1)	85.3 (83.2, 87.2)	-	79.2 (76.7, 81.4)	64.2 (61.4, 67.0)	86.7 (84.6, 88.6)	-

cancer registry. When recorded in multiple datasets, cancer dates were generally earlier in cancer registry data compared to primary care data, but on the same date or later in cancer registry compared to secondary care data. Agreement in recording of case status does not appear to have increased over time.

4.2. *Strengths and limitations*

Within each data source, diagnoses were determined from the presence of a single code, and were not confirmed using other codes, investigations, or treatments. The HES Outpatient (OP) dataset was not included in this study, as the recording of diagnostic fields is not mandatory and has high missingness [21,22]. Additionally, HES only contains data on NHS funded hospital care – diagnoses in private healthcare settings may not be captured.

To maximise the chances of cancer registry diagnoses being captured in primary and/or secondary care, records of each cancer site prior to a patient's registration at an up-to-standard CPRD GOLD practice were discarded from all data sources, in line with a previous study [10]. Thus, for some cases, the earliest record of their cancer diagnosis in HES or CR may have been ignored. However, as a year of disease-free follow-up prior to diagnosis was required in each data source, it is unlikely that these excluded records represent diagnoses that would have been captured in CPRD.

As previous studies have shown, agreement in recorded cancer diagnosis is likely to vary between cancer sites [8,10]. As an initial approach, we included five distinct cancer sites as it was not possible to consider all cancer sites in this study. The included sites have different presenting signs and symptoms, and appreciable proportions of cases with these cancers are diagnosed through different diagnostic routes (for example, emergency presentations, two-week wait referral pathways, or screening) which may impact the recording and timing of cancer diagnosis in data sources differently [23,24]. However, the availability and use of different pathways has varied over time and diagnostic pathway variation is not the only reason why concordance between data sources may vary by cancer site. Future research should expand this analysis to other cancers such as non-melanoma skin cancers, rectal cancers, and haematological cancers for which concordance may vary for other reasons – for instance GPs may be more likely to record suspected diagnoses of non-melanoma skin cancers. Furthermore, for patients with a cancer diagnosis recorded in CR, future research could examine the extent to which diagnostic route is associated with a diagnosis also being recorded in other data sources.

Brain, pancreatic, and lung cancer have relatively high proportions of patients diagnosed as an emergency [24,25] and can have high mortality rates [26,27]. We selected cases from a cohort with at least one year of follow-up between 2007 and 2021. As fewer patients diagnosed between 1999 and 2007 will have survived for long enough to be eligible for inclusion these sites are underrepresented prior to 2008. It is unclear how this affects results, however the high rate of emergency presentations may explain the high probability of a CR diagnosis also being recorded in HES for these sites (Supplementary Figures 6, 8 & 10).

When identifying brain tumour diagnoses, both benign and malignant diagnoses were included – as common in cancer registration practice – as the presenting symptoms and diagnostic processes are similar. Uniquely among the cancer sites studied, a large proportion of brain tumour cases identified in CPRD and HES did not have their diagnosis recorded in cancer registry data. There are two theoretical explanations for this finding.

First, that some brain cancer cases recorded in data sources other than the cancer registry represent 'false positives'. This can relate to situations where there was initial suspicion of brain cancer, which was not subsequently confirmed; or instances where brain metastases originating from primary tumours of other organs are misclassified as brain tumours. Brain is a very common site of metastasis from primary tumours in other organs, so the ratio of secondary to primary tumours in

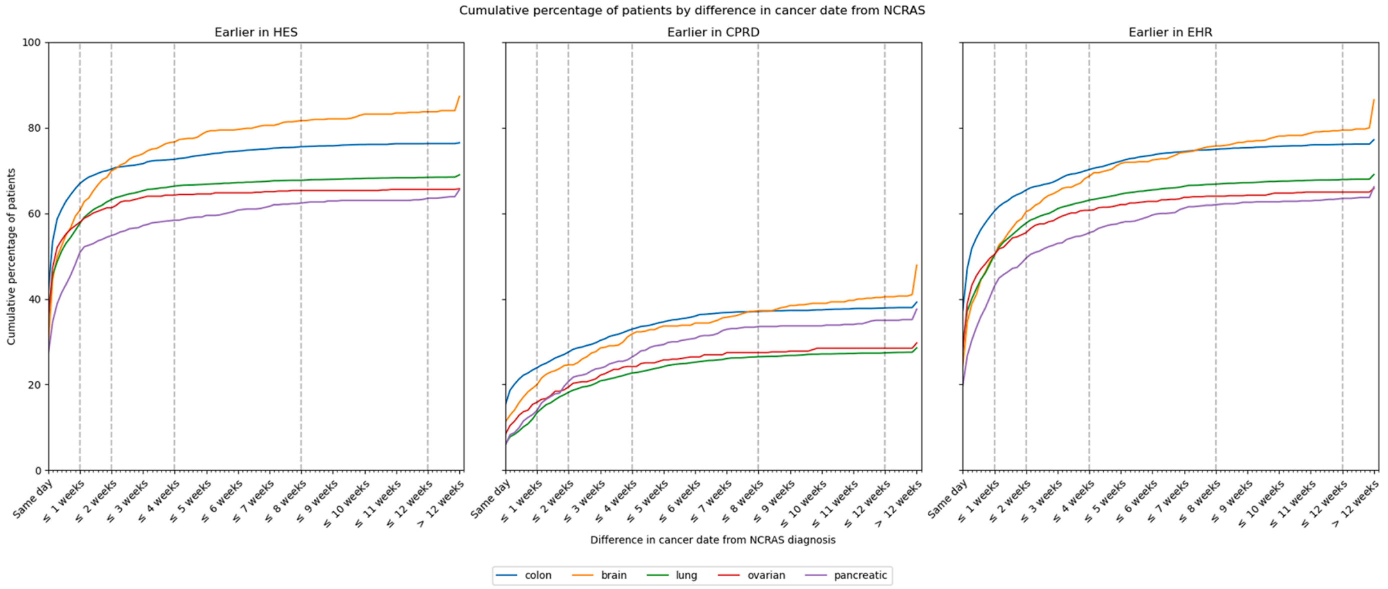


Fig. 1. Cumulative percentage of patients by difference in cancer date (in a secondary data source - HES, CPRD, or combined EHR) from NCRAS cancer date, for diagnoses occurring earlier in a secondary data source than in NCRAS. For each data source, the denominator population includes patients identified as having a cancer diagnosis recorded in both NCRAS (2000 – 2017) and another data source (1999 – 2018) within 1 year.

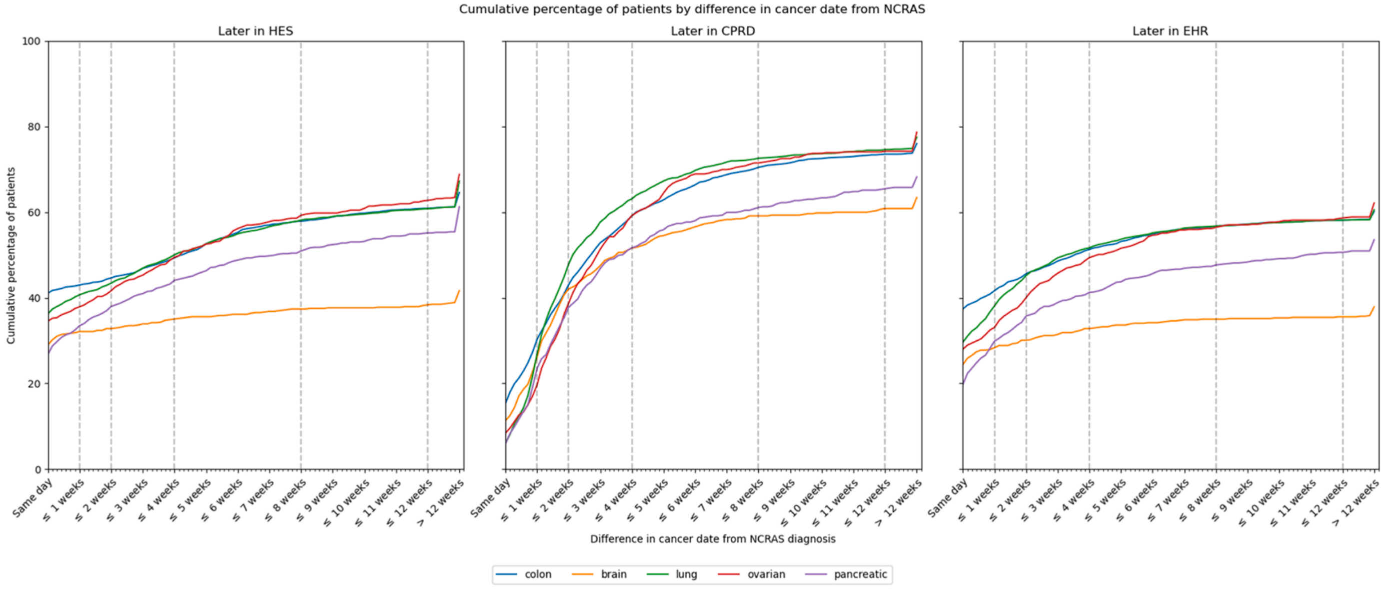


Fig. 2. Cumulative percentage of patients by difference in cancer date (in a secondary data source - HES, CPRD, or combined EHR) from NCRAS cancer date, for diagnoses occurring later in a secondary data source than in NCRAS. For each data source, the denominator population includes patients identified as having a cancer diagnosis recorded in both NCRAS (2000 – 2017) and another data source (1999 – 2018) within 1 year.

the brain is high. Therefore, a small degree of misclassification of metastases could produce a large degree of discordance between different sources, as observed in our study. Examining recording concordance between data sources for neoplasms of other organs where the majority of neoplasms arise from other organs, such as liver, would be useful.

Second, some brain cancers not recorded by the cancer registry may represent 'false negatives', i.e., genuine primary brain tumours that have not been captured. Further research is needed to elucidate the relatively high degree of discordance observed for brain tumours.

4.3. Comparison with literature

There are seven main papers of relevance to this study [6–10,28,29] – summarised in Supplementary Information (Supp. Table 14) – all using data from prior to 2014. Key comparisons are outlined in Table 3. Studies showed wide ranges in agreement between primary and secondary care records and cancer registry, with earlier studies appearing to show higher concordance [9,28]. Our study is the first to examine whether concordance between CR, CPRD GOLD, and HES has changed during a recent 20-year period.

Agreement between CR and CPRD is more variable over time and across cancer sites compared to agreement between CR and HES; this is likely to reflect that a range of factors may impact the recording of cancer diagnosis in primary care. Given the increased use of computers during primary care consultations, and changes to the design of general practice EHR systems, more recent coded data may be more accurate

Table 3
Key comparisons with existing literature.

Key observations from the literature	Comparison with results of this study
More CR cases were recorded in HES than CPRD (although less likely for cancers managed in primary care) but using a combination of datasets was typically more sensitive than a single dataset [6,8,10].	Our findings align with those from previous studies, with HES records confirming more CR diagnoses than CPRD. A combined EHR dataset (CPRD + HES) was able to identify a higher percentage of CR cases than either CPRD or HES alone for all five cancer sites (83.9 – 92.0 %).
Diagnoses in CPRD were likely to occur later than in CR [6,9] – Boggon et al. noted that 63 % of regional cancer registry cases were also captured in CPRD within 1 month [7].	This aligns with our finding that CPRD diagnoses lag behind CR cancer dates, but that between 72.1 – 79.8 % of CR cases have the diagnosis captured in CPRD within 1 month.
HES shows the highest concordance in cancer dates with CR [6]	This was also the case in our study – between 26.8 – 41.1 % of CR patients had matching cancer dates in HES, compared to 5.8 – 15.3 % in CPRD and 19.7 – 37.4 % in combined EHR data.
Margulis et al. observed greater completeness of CPRD records between 2004 and 2008 [8]. Boggon et al. and Margulis et al. were able to validate some CPRD cases using free-text from medical records [7,8].	Supplementary figures 6 - 10 show the probability of a CR diagnosis being recorded in CPRD peaking during this period for three cancer sites (colon, lung, ovary) – possibly reflecting an initial improvement in recording of cancer diagnosis in primary care following the introduction of the Quality and Outcomes Framework in 2004 [30]. We were unable to access free-text to validate CPRD cases.

and complete than before. Additionally, the introduction of the ’two-week wait’ strategy since the 2000’s, and Quality and Outcomes Framework in 2004 may have affected the recording of cancer diagnosis in primary care [30–33]. These factors – combined with changes to cancer registration – could explain some of the heterogeneity between CPRD and CR recording over time.

4.4. Implications

The datasets compared in this study are collected for different purposes and, to some extent, cancer dates may reasonably differ between them. Myklebust et al. illuminate how the date of incidence used in CR varies with cancer registration practices [29]. When registration is based on histological confirmation (as in England, per ENCR rules [15]) this ‘delays’ diagnosis from the date of first relevant hospital admission or clinical encounter. The resulting NCRAS cancer date is likely to occur on the same day or after the HES APC cancer date (Figs. 1 and 2).

Comparing cancer dates between primary care and cancer registration is more complex. Primary care cancer dates depend on recording practices of individual clinicians and wider healthcare system factors, such as changing cancer detection strategies. Some clinicians may code a *suspicion* of cancer using diagnostic codes – in which case CPRD cancer dates will appear before CR dates. Others may record cancer only after formal diagnosis – for example, on receipt of confirmation from secondary care – or may backdate the cancer date to a hospital admission or pathology date. This could result in the heterogeneous agreement we observe between CR and CPRD (Figs. 1 and 2).

It is also worth considering whether data from primary care could improve the quality and coverage of the cancer registry. Whilst data streams from secondary care are already incorporated into the cancer registration process, data does not currently pass directly from primary care settings to the cancer registry [14]. As described above, the diagnoses coded in primary care are likely too tenuous to be relied upon for case ascertainment purposes and are not used to determine incidence dates. However, primary care data could be informative for the ascertainment of other covariates, such as comorbidity burden.

The optimal dataset(s) for determining cancer dates depends on the research question. For example, for diagnostic quality and safety research [34], using cancer registry data to determine cancer date may introduce complexity and bias. The date of histological confirmation recorded in cancer registry data (in England) may differ from the date the diagnosis was communicated to the patient based on other investigations (such as imaging). Other datasets may be preferable for determining this diagnosis date [34]. It is unclear how study designs that depend on a *diagnosis* date may be biased by differences in the cancer date, however variation in case status alone may be sufficient to bias estimates of epidemiological measures of cancer outcomes. Possible implications of differences in case status and cancer date on different study designs are outlined in Box 1.

Similarly the codes used to record the same tumour may vary between data sources. As illustrated in Supplementary Table 7, this issue affects all cancer sites to varying degrees, but the largest variations can be seen for ovarian and colon cancers. For instance, assigning a site to ovarian cancer can be challenging and tumours that may be recorded in HES as ovarian (i.e., ICD10 code C56) may be assigned other sites, such as peritoneum, in the cancer registry. Note that guidance from the Royal College of Pathologists advises pathologists to assign a site of fallopian tube or primary peritoneum to what may be clinically referred to as ‘ovarian cancer’ based on the distribution of disease seen microscopically [38]. The cancer registry may be more likely to record the pathologist assigned site than the HES assigned site. As such, the accuracy of site coding should be considered when selecting datasets for research and when determining phenotypes.

4.5. Recommendations for selecting datasets

Costs, access delays, and linkage availability of cancer registration data may present barriers to research [1]. Our findings suggest that using combined electronic health record data for research can reasonably approximate cancer registry data, although limitations – such as variable agreement across cancer sites and over time – must be considered and mitigated where possible.

Researchers should note that 84 – 92 % of cancer registry cases were recorded in combined EHR data, and, except for brain cancer, similar percentages of combined EHR cases were recorded in the cancer registry; this highlights that agreement is high, but not perfect, and for brain cancer it is indeed poorer. Clinician record review has previously confirmed that many cancer diagnoses recorded in primary care are valid [8], though research is needed to determine whether this remains true. Recommendations for using different data sources to determine case status and cancer date are summarised in Table 4. When selecting data sources, researchers should consider the purpose for which the data was collected, the data collection methods, and the possible biases caused by variation in case status and diagnosis date in the context of their specific research question.

5. Conclusion

Cancer registration data are considered the reference standard for determining cancer incidence, but differences exist between case status and cancer date compared to electronic health records in a minority of patients with four of the five cancers studied (*colon, lung, ovary, and pancreas*), while more substantial differences exist for brain cancer. Combined data from primary and secondary care data may be sufficient to identify cases in many circumstances. Further, while the CR is the reference standard for fact of cancer, the earliest record of cancer in the electronic health record may be the most relevant date for research into the clinical diagnosis of cancer. The date we consider to represent the diagnosis may impact on diagnostic interval lengths, positive predictive values of presenting symptoms, and the distribution of diagnostic routes. Mechanisms responsible for disagreements in cancer diagnosis status and its recorded timing between sources should be elucidated by future research, to help guide choice of the most appropriate diagnosis date.

Box 1
Possible implications of differences in case status and cancer date on study designs dependent on a diagnosis date.

Implications for studies looking back from the diagnosis date

1. ***Diagnostic windows*** – Differences in case status and cancer date may impact estimates of diagnostic window length, dependent on the underlying reasons for and directionality of differences.
2. ***Prodromal features*** – Many symptom-based NICE guidelines are based on research using primary care records to determine diagnosis. As both the cases identified and cancer dates differ between data sources, this could affect PPV estimates.
3. ***Diagnostic routes*** – Differences in case status may lead to differing proportions of routes in the population both through 'direct' misclassification of incident cases and indirectly as differences in cancer date may misclassify 'emergency presentation' status (typically defined by emergency healthcare use in the 30 days prior to the diagnosis date).
4. ***Diagnostic intervals*** – Measurement of intervals relies on the identification of a start point (typically a symptomatic consultation) and an end point (the diagnosis date). As such, variation in the cancer date will result in different estimates of diagnostic intervals in CR compared to EHR data. Differences in case status may be driven by factors that also relate to the length of the interval.
5. ***Missed diagnostic opportunities*** – Captured at the individual-level, differences in cancer date will lead to different amounts of time for potential missed diagnostic opportunities to have occurred in different datasets, possibly biasing results.

Implications for studies looking forwards from the diagnosis date

6. ***Survival and prognostic outcomes*** – Differences in cancer date may result in immortal time bias. Differences in case status may be driven by factors that also relate to survival (e.g., death-certificate only diagnoses may only be captured in cancer registration data)
7. ***Patient surveys*** – Different cancer patient surveys use different methods to draw patient samples [35–37]. Differences in cancer date and case status between datasets may result in differences in case-mix.

Implications for studies dependent on case status

8. ***Incidence studies –*** Differences in case status will likely result in variation in estimated incidence levels. Factors driving differences in case status may cause over- or under-estimation of incidence levels in certain patient groups.

Table 4
Implications and recommendations for using individual data sources to determine case status and diagnosis date.

Available data source	Implications	Recommendations
Primary care data	Wide variation likely in recording of cancer diagnoses – consider case status and cancer date with caution False positives and false negatives are possible	Determining cancer status and/or diagnosis date solely from primary care data should be avoided when possible - confirmation using a second data source could help limit false negatives. If necessary, a higher threshold should be considered to limit false positives – e.g., using multiple records of the condition, or evidence of treatment.
Hospital admission data	Case status and cancer date likely more robust than primary care data Some cases – particularly those managed in primary care, such as skin cancer in some health systems [39] – are likely to be missing [8]	Consider diagnostic processes and treatment pathways in the healthcare system in question to identify whether any particular cancer sites or patient groups may be missing
Population-based cancer registration data	Case status likely to be accurate. Cancer date should be consistent with registration practices.	The ability to interpret the date of incidence recorded in registration data as a proxy for 'date of diagnosis' will be dependent on specific registration practices.

Funding information

EW acknowledges the receipt of a studentship award from the Health Data Research UK-The Alan Turing Institute Wellcome PhD Programme in Health Data Science (Grant Ref: 218529/Z/19/Z). BW, MEB, and GL were supported by a Cancer Research UK (C18081/A18180) Advanced Clinician Scientist Fellowship (to GL). MEB is supported by a Cancer Research UK International Alliance for Cancer Early Detection (ACED) Pathway Award (EDDAPA-2022/100002). CR acknowledges funding from Cancer Research UK — Early Detection and Diagnosis Committee (EDDCPJT/100018). The study aligns to (but was not directly supported by) the RREDD-EHR project supported by the International Alliance for Cancer Early Detection a partnership between Cancer Research UK (C18081/A31373), Canary Center at Stanford University, the University of Cambridge, OHSU Knight Cancer Institute, University College London, and the University of Manchester. SD is funded from: (a) Health Data Research UK, which receives its funding from HDR UK Ltd (HDR-9006) funded by the UK Medical Research Council, Engineering and Physical Sciences Research Council, Economic and Social Research Council, Department of Health and Social Care (England), Chief Scientist Office of the Scottish Government Health and Social Care Directorates, Health and Social Care Research and Development Division (Welsh Government), Public Health Agency (Northern Ireland), British Heart Foundation (BHF) and the Wellcome Trust. (b) the BigData@Heart Consortium, funded by the Innovative Medicines Initiative-2 Joint Undertaking under Grant agreement No 116074., (c) the NIHR UCLH Biomedical Research Centre, (d) an Alan Turing Fellowship (EP/N510129/1), (e) the Longitudinal Health and Wellbeing COVID-19 National Core Study, which was established by the UK Chief Scientific Officer in October 2020 and funded by UK Research and Innovation (Grant references MC_PC_20030 and MC_PC_20059), by the Data and Connectivity National Core Study, led by Health Data Research UK in partnership with the Office for National Statistics and funded by UK Research and Innovation (Grant reference MC_PC_20058), and by the CONVALESCENCE study of long COVID, which is funded by NIHR/UKRI, (f) The British Heart Foundation Data Science Centre (Grant No SP/19/3/34678, awarded to Health Data Research (HDR) UK).

Data Statement

This study protocol was approved by the UK Medicines and

Healthcare products Regulatory Agency (MHRA) Independent Scientific Advisory Committee (ISAC Protocol number 18_299), under Section 251 (NHS Social Care Act 2006). This study is based on data from the CPRD obtained under licence from the MHRA. The data are provided by patients and collected by the NHS as part of their care and support. The interpretation and conclusions contained in this study are those of the authors alone.

Codelists used in this study are available online at https://github.com/ekw26/CR-EHR-phenotypes

CRediT authorship contribution statement

Spiros Denaxas: Writing – review & editing, Supervision. **Georgios Lyratzopoulos:** Writing – review & editing, Supervision, Conceptualization. **Cristina Renzi:** Writing – review & editing. **Brian Rous:** Writing – review & editing, Conceptualization. **Emma Whitfield:** Writing – original draft, Software, Methodology, Formal analysis, Conceptualization. **Matthew E Barclay:** Writing – review & editing, Supervision, Methodology. **Meena Rafiq:** Writing – review & editing. **Becky White:** Writing – review & editing, Supervision, Methodology.

Declaration of Competing Interest

MEB receives personal fees from GRAIL Inc., for Independent Data Monitoring Committee (IDMC) membership unrelated to this study. All other authors declare no competing interests.

Appendix A. Supporting information

Supplementary data associated with this article can be found in the online version at doi:10.1016/j.canep.2024.102703.

References

[1] NHS England, Correspondence - Freedom of Information Request 2307-2006940, NHSE:0141511, (2023).

[2] A.H. Siddiqui, S.N. Zafar, Global availability of cancer registry data, J. Glob. Oncol. (4) (2018), https://doi.org/10.1200/JGO.18.00116.

[3] T.P.C. Chu, A. Shah, D. Walker, M.P. Coleman, Pattern of symptoms and signs of primary intracranial tumours in children and young adults: a record linkage study, Arch. Dis. Child 100 (2015) 1115–1122, https://doi.org/10.1136/ARCHDISCHILD-2014-307578.

[4] N.L. Barclay, M. Pineda Moncusí, A.M. Jödicke, D. Prieto-Alhambra, B. Raventós, D. Newby, A. Delmestri, W.Y. Man, X. Chen, M. Català, The impact of the UK COVID-19 lockdown on the screening, diagnostics and incidence of breast, colorectal, lung and prostate cancer in the UK: a population-based cohort study, Front. Oncol. 14 (2024), https://doi.org/10.3389/FONC.2024.1370862.

[5] T.A. Ahmad, A.Z.D. Ullah, C. Chelala, D.P. Gopal, F. Eto, R. Henkin, M. Samuel, S. Finer, S.J. Taylor, Prevalence of multimorbidity in survivors of 28 cancer sites: an English nationwide cross-sectional study, Am. J. Cancer Res. 14 (2024) 880, https://doi.org/10.62347/NWHM4133.

[6] C.S. Arhi, A. Bottle, E.M. Burns, J.M. Clarke, P. Aylin, P. Ziprin, A. Darzi, Comparison of cancer diagnosis recording between the clinical practice research datalink, cancer registry and hospital episodes statistics, Cancer Epidemiol. 57 (2018) 148–157, https://doi.org/10.1016/J.CANEP.2018.08.009.

[7] R. Boggon, T.P. Van Staa, M. Chapman, A.M. Gallagher, T.A. Hammad, M. A. Richards, Cancer recording and mortality in the General Practice Research Database and linked cancer registries, Pharmacoepidemiol Drug Saf. 22 (2013) 168–175, https://doi.org/10.1002/PDS.3374.

[8] A.V. Margulis, J. Fortuny, J.A. Kaye, B. Calingaert, M. Reynolds, E. Plana, L. J. McQuay, W.J. Atsma, B. Franks, S. De Vogel, S. Perez-Gutthann, A. Arana, Validation of cancer cases using primary care, cancer registry, and hospitalization data in the United Kingdom, Epidemiology 29 (2018) 308, https://doi.org/10.1097/EDE.0000000000000786.

[9] A. Dregan, H. Moller, T. Murray-Thomas, M.C. Gulliford, Validity of cancer diagnosis in a primary care database compared with linked cancer registrations in England. Population-based cohort study, Cancer Epidemiol. 36 (2012) 425–429, https://doi.org/10.1016/J.CANEP.2012.05.013.

[10] H. Strongman, R. Williams, K. Bhaskaran, 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, BMJ Open 10 (2020) e037719, https://doi.org/10.1136/BMJOPEN-2020-037719.

[11] E. Herrett, A.M. Gallagher, K. Bhaskaran, H. Forbes, R. Mathur, T. van Staa, L. Smeeth, Data resource profile: clinical practice research datalink (CPRD), Int. J. Epidemiol. 44 (2015) 827, https://doi.org/10.1093/IJE/DYV098.

[12] Medicines & Healthcare Products Regulatory Agency (MHRA), Clinical Practice Research Datalink (CPRD), Hospital Episode Statistics (HES) Admitted Patient Care and CPRD primary care data Documentation (set 22/January 2022), (2022).

[13] Public Health England (PHE), The National Cancer Registration and Analysis Service A guide to cancer data and working with us, 2020.

[14] K.E. Henson, L. Elliss-Brookes, V.H. Coupland, E. Payne, S. Vernon, B. Rous, J. Rashbass, Data resource profile: national cancer registration dataset in England, Int. J. Epidemiol. 49 (2020) 16, https://doi.org/10.1093/IJE/DYZ076.

[15] European Network of Cancer Registries, ENCR Recommendations: Coding Incidence Date (2022), 2022. https://encr.eu/sites/default/files/Recommendations/ENCR%20Recommendation%20DOI_Mar2022_0.pdf (accessed July 13, 2023).

[16] Medicines & Healthcare products Regulatory Agency (MHRA), Clinical Practice Research Datalink (CPRD), The Public Health England National Cancer Registration and Analysis Service (NCRAS) and CPRD primary care data Documentation (set 21), (2021).

[17] E. Whitfield, CR-EHR-phenotypes: Published version of codelists, (2024). https://doi.org/10.5281/ZENODO.13710791.

[18] A. Elixhauser, C. Steiner, D.R. Harris, R.M. Coffey, Comorbidity measures for use with administrative data, Med. Care 36 (1998) 8–27, https://doi.org/10.1097/00005650-199801000-00004.

[19] H. Quan, V. Sundararajan, P. Halfon, A. Fong, B. Burnand, J.C. Luthi, L. D. Saunders, C.A. Beck, T.E. Feasby, W.A. Ghali, Coding algorithms for defining comorbidities in ICD-9-CM and ICD-10 administrative data, Med. Care 43 (2005) 1130–1139, https://doi.org/10.1097/01.MLR.0000182534.19832.83.

[20] Cancer Outcomes and Services Data set (COSD) - NDRS, (n.d.). https://digital.nhs.uk/ndrs/data/data-sets/cosd (accessed October 8, 2024).

[21] Medicines & Healthcare Products Regulatory Agency (MHRA), Clinical Practice Research Datalink (CPRD), Hospital Episode Statistics (HES) Outpatient Care and CPRD primary care data Documentation (set 21/ August 2021), (2021). https://doi.org/10.48329/cp5e-7790.

[22] A. Boyd, R. Cornish, L. Johnson, S. Simmonds, H. Syddall, L. Westbury, C. Cooper, J. Macleod, Understanding Hospital Episode Statistics (HES), London, 2017. https://www.closer.ac.uk/wp-content/uploads/ CLOSER-resource-understanding-hospital-episode-statistics-2018.pdf (accessed November 8, 2022).

[23] L. Elliss-Brookes, S. McPhail, A. Ives, M. Greenslade, J. Shelton, S. Hiom, M. Richards, Routes to diagnosis for cancer – determining the patient journey using multiple routine data sets, Br. J. Cancer 107 (2012) 1220–1226, https://doi.org/10.1038/bjc.2012.408.

[24] G.A. Abel, J. Shelton, S. Johnson, L. Elliss-Brookes, G. Lyratzopoulos, Cancer-specific variation in emergency presentation by sex, age and deprivation across 27 common and rarer cancers, Br. J. Cancer 112 (2015) S129, https://doi.org/10.1038/BJC.2015.52.

[25] S. McPhail, R. Swann, S.A. Johnson, M.E. Barclay, H. Abd Elkader, R. Alvi, A. Barisic, O. Bucher, G.R.C. Clark, N. Creighton, B. Danckert, C.A. Denny, D. W. Donnelly, J.J. Dowden, N. Finn, C.R. Fox, S. Fung, A.T. Gavin, E. Gomez Navas, S. Habbous, J. Han, D.W. Huws, C.G.C.A. Jackson, H. Jensen, B. Kaposhi, S. E. Kumar, A.L. Little, S. Lu, C.A. McClure, B. Møller, G. Musto, Y. Nilssen, N. Saint-Jacques, S. Sarker, L. te Marvelde, R.S. Thomas, R.J.S. Thomas, C.S. Thomson, R. R. Woods, B. Zhang, G. Lyratzopoulos, B. Filsinger, K. Forster, L. May, D. S. Morrison, A.F. Thomas, J.L. Warlow, H. You, Risk factors and prognostic implications of diagnosis of cancer within 30 days after an emergency hospital admission (emergency presentation): an International Cancer Benchmarking Partnership (ICBP) population-based study, Lancet Oncol. 23 (2022) 587–600, https://doi.org/10.1016/S1470-2045(22)00127-9.

[26] Cancer Research UK, Lung cancer mortality statistics, (n.d.). https://www.cancerresearchuk.org/health-professional/cancer-statistics/statistics-by-cancer-type/lung-cancer/mortality (accessed July 28, 2023).

[27] Cancer Research UK, Pancreatic cancer mortality statistics, (n.d.). https://www.cancerresearchuk.org/health-professional/cancer-statistics/statistics-by-cancer-type/pancreatic-cancer/mortality (accessed July 28, 2023).

[28] H. Møller, S. Richards, N. Hanchett, S.P. Riaz, M. Lüchtenborg, L. Holmberg, D. Robinson, Completeness of case ascertainment and survival time error in English cancer registries: impact on 1-year survival estimates, Br. J. Cancer 105 (2011) 170–176, https://doi.org/10.1038/bjc.2011.168.

[29] T.Å. Myklebust, T. Andersson, A. Bardot, S. Vernon, A. Gavin, D. Fitzpatrick, M. B. Jerm, M. Rutherford, D.M. Parkin, P. Sasieni, M. Arnold, I. Soerjomataram, F. Bray, P.C. Lambert, B. Møller, Can different definitions of date of cancer incidence explain observed international variation in cancer survival? An ICBP SURVMARK-2 study, Cancer Epidemiol. 67 (2020) 101759, https://doi.org/10.1016/J.CANEP.2020.101759.

[30] National Health Service (NHS), Quality and Outcomes Framework (QOF) 2004/05 background, 2012.

[31] National Health Service (NHS), The NHS Cancer Plan, 2000. https://webarchive.nationalarchives.gov.uk/ukgwa/20130107105354/http://www.dh.gov.uk/prod_consum_dh/groups/dh_digitalassets/@dh/@en/documents/digitalasset/dh_4014513.pdf (accessed September 11, 2023).

[32] NHS Improvement, Department of Health, National Cancer Action Team, National Cancer Intelligence Network, Ensuring Better Treatment: Going Further on Cancer Waits, 2008.

[33] J.S. Taggar, T. Coleman, S. Lewis, L. Szatkowski, The impact of the Quality and Outcomes Framework (QOF) on the recording of smoking targets in primary care medical records: cross-sectional analyses from The Health Improvement Network (THIN) database, BMC Public Health 12 (2012) 329, https://doi.org/10.1186/1471-2458-12-329.

[34] Committee on Diagnostic Error in Health Care, Board on Health Care Services, Institute of Medicine, The National Academies of Sciences Engineering and Medicine, Improving Diagnosis in Health Care, National Academies Press (US), 2015. https://doi.org/10.17226/21794.

[35] Picker Institute, National Cancer Patient Experience Survey, (n.d.). https://www.ncpes.co.uk/ (accessed December 14, 2023).

[36] Picker Institute, National Cancer Patient Experience Survey Programme Sampling Instructions 2023, (2023).

[37] National Health Service (NHS) England, Cancer Quality of Life Survey, (n.d.). https://www.cancerqol.england.nhs.uk/ (accessed December 14, 2023).

[38] W.G. McCluggage, M.J. Judge, B.A. Clarke, B. Davidson, C.B. Gilks, H. Hollema, J. A. Ledermann, X. Matias-Guiu, Y. Mikami, C.J.R. Stewart, R. Vang, L. Hirschowitz, Data set for reporting of ovary, fallopian tube and primary peritoneal carcinoma: recommendations from the International Collaboration on Cancer Reporting (ICCR), Mod. Pathol. 2015 28:8 28 (2015) 1101–1122, https://doi.org/10.1038/modpathol.2015.77.

[39] C.B. Del Mar, J.B. Lowe, The skin cancer workload in Australian general practice, Aust. Fam. Physician 26 (1) (1997) S24–S27.