Rapid Cancer Registration Dataset: data at 4th February (CAS2302)

The National Cancer Registration and Analysis Service (NCRAS) has developed an algorithmically generated Rapid Cancer Registration Dataset (RCRD) using the standard administrative datasets which flow rapidly into NHS England (NHSD) and are incorporated into the Cancer Analysis System (CAS) of NCRAS. The data takes the form of a series of significant events that occur to each patient as they proceed through the diagnostic and then therapeutic parts of the cancer pathway, and is available at approximately 4-5 months behind real time. The RCRD is shallower and narrower than the full NCRAS cancer registration dataset; it should be used and interpreted with reference to the caveats outlined within this document.

Main findings

This document outlines the main features of the data to be aware of when interpreting the Rapid Cancer Registration Dataset:

- Across all cancers types included approximately 11.5% of cases are missing and 6.1% of cases are included erroneously or with incorrect cancer type or diagnosis date (when compared to 'Gold Standard' registration data for 2018 data).
- These figures vary strongly with cancer site. Broadly, more common cancers (particularly breast and prostate cancer) perform best and less common cancers (particularly bone and soft tissue and cancers of unknown primary) perform worst.
- Non-melanoma skin cancer (ICD-10 C44) tumours are excluded from the majority of data shown (Figure 3 onwards). Carcinoma of the cervic (ICD-10 D06) is excluded from all data presented.
- There are more missing tumours in those aged over 70 compared to younger age groups.
- Other factors that reduce data completeness include the patient's route to diagnosis, mortality within 30 days or diagnosis, and the presence of multiple cancers.
- Usable data is available approximately 4-5 months after diagnosis or other clinical activity occurs.
- Data on cancer stage group at diagnosis is available for a number of common tumour types, although completeness is lower than that for the Gold Standard registration data. Where data is available it generally agrees with the Gold Standard stage group in 80-90% of tumours.

The dataset includes Rapid Cancer Registrations from January 2018 to the most recently available data (at the date specified in the title to this document), plus additional event data for the same period.

Contents

Summary

A need to make rapidly available 'proxy cancer registrations' (and associated clinical activity) for the COVID-19 period has been identified to support the public health response by NHS England (PHE) and other agencies, and service reorganisation by the NHS. These proxy registrations are called Rapid Registrations in contrast to the more formal detailed registration process that are used in non-clinical cancer research and the National Statistics (https://www.gov.uk/government/statistics/cancer-registration-statistics-england-2018-final-release).

The National Cancer Registration and Analysis Service (NCRAS) has developed a Rapid Cancer Registration Dataset (RCRD) using all standard administrative datasets which flow rapidly into PHE and are incorporated into the Cancer Analysis System (CAS) of NCRAS.

This document describes the dataset structure, creation methodology, and data quality caveats (due to the rapid automated creation process without additional data curation) behind this dataset.

These data structures and methodologies are expected to evolve over the course of the public health response to COVID-19. The data is updated monthly and is referred to by the monthly CAS snapshot upon which it is based, e.g. CAS2009 refers to the CAS snapshot from September 2020. This document is considered a 'living document' and strictly applies only to the snapshot of CAS identified in the title.

Methodology

Proxy registration events (Rapid Registrations)

Datasets available to PHE were surveyed for how many months in arrears that they arrive within NCRAS and are loaded in a usable format for analysis. From these datasets a selection of event types were defined similarly to those typically used for cancer pathway analysis pursued by NCRAS.

The data takes the form of a series of significant events that occur to each patient as they proceed through the diagnostic and then therapeutic parts of the cancer pathway. These events include chemotherapy cycles, radiotherapy episodes and major cancer surgery as well as events based

on the Cancer Waiting Times (CWT) and Cancer Outcomes and Services Dataset (COSD) datasets. These event types are numbered in the range 1-23 in the dataset.

Some events hypothesised to be indicative of a cancer diagnosis were defined including 'Diagnosis reported in COSD' (event 51) and 'CWT estimated diagnosis date' (event 52). These are numbered in the range 50-57 in the dataset - see Appendix 1 for a full list.

The indicative events for diagnosis were explored as candidate Rapid Registration events. These candidate rapid registration events were judged as matching against a Gold Standard Registration event if it met the following two conditions:

- The difference in diagnosis dates for each event was 90 days or less.
- Both registrations fell into the same broad tumour group (as defined in Appendix 3).

Using these matching criteria False Positive errors and False Negative errors are defined as:

- **False Positive Error (FPE)**: A rapid registration event has been created which does not match against a Gold Standard Registration in the comparison period.
- **False Negative Error (FNE)**: There exists a Gold Standard Registration event for which no rapid registration event can be matched.

Additional filtering was applied to the candidate events and eventually event 101 was defined to minimise both false positive and false negative errors and is recommended for use by researchers as the best candidate for a rapid cancer registration. Appendix 4 briefly examines some of the alternatives examined in the development of this event definition.

Data structures

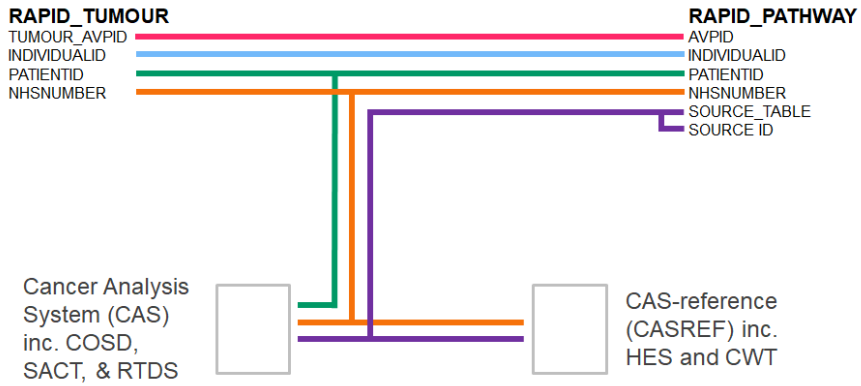
The rapid registration dataset consists of two tables:

AT_RAPID_PATHWAY: This is an event-based dataset with a number of types of event of interest defined based on the rapidly available datasets, see Appendix 1 for event definitions and properties. These are numbered in the range 1-23 for general purpose events, 50-57 for events that are candidates for combining into a rapid registration, and 101 for the final rapid registration event.

AT_RAPID_TUMOUR: This is a tumour level dataset that holds tumour and patient level data for each of the tumours defined by a rapid registration. The structure and contents of this table are presented in Appendix 3.

The rapid registration pathway and tumour table can be linked together as shown in Figure 1, and also to other datasets that are timely enough via NHSnumber.

Figure 1: Linkage diagram for the Rapid Cancer Registration Dataset



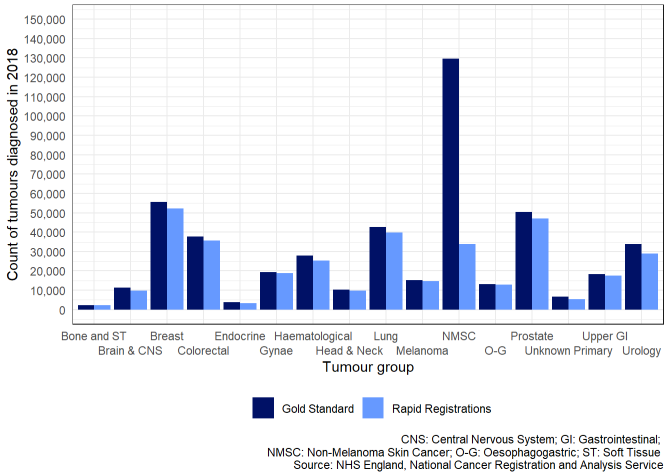
Data Quality

How do the number of Rapid Registrations compare with Gold Standard Registrations?

To illustrate the strengths and weaknesses of the Rapid Registrations compared to the gold standard process, registrations for tumours diagnosed during 2018 are compared in Figure 2.

For most tumour groups the counts of Rapid Registrations are significantly lower than those of standard registrations. The COSD system does not attempt to record basal cell carcinoma non-melanoma skin cancers (but they are recorded by hospital pathology systems, and thereby registered), explaining the discrepancy there. There is only one group where this situation is reversed - bone and soft tissue - for which a precise morphology is required to properly record the diagnosis. These cancers are being preferentially coded to bone and soft tissue in COSD (as the COSD standard necessitates simpler site-based coding, and this is the best choice under the circumstances) and re-coded during the gold standard registration process where more sophisticated combination of site and morphological coding is possible.

Figure 2: The number of cancer registrations by registration and tumour type, England, 2018

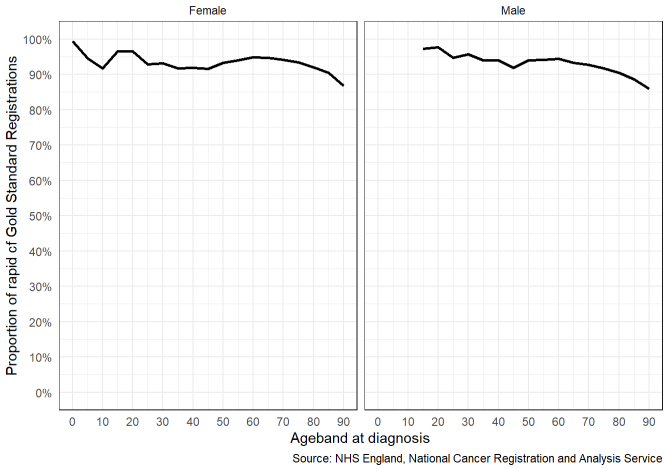


CNS: Central Nervous System; GI: Gastrointestinal;
NMSC: Non-Melanoma Skin Cancer; O-G: Oesophagogastric; ST: Soft Tissue
Source: NHS England, National Cancer Registration and Analysis Service

Figure 3 shows the age dependence of the ratio between Gold Standard and Rapid Registrations, Non-Melanoma Skin Cancer is excluded. The

proportion of diagnoses is consistently high for both males and females until the age of 70 is reached, where it declines. This is explored further in Figure 5 below.

Figure 3: The proportion of cancer registrations by sex, age and registration type, England, 2018 (all tumour types combined)



Source: NHS England, National Cancer Registration and Analysis Service

Comparing the matching quality of Rapid Registrations

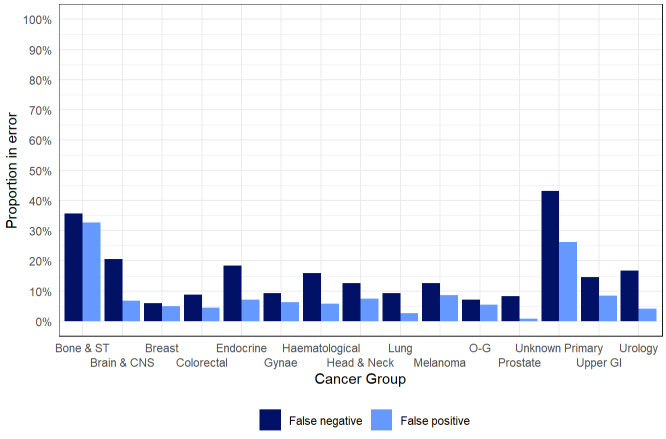
The quality of the Rapid Registrations was judged by comparing them against the gold-standard cancer registrations in the period April 2018 to September 2018. This period was chosen as available gold standard registration data was only finalised to December 2018 and a matching period of 90 days was allowed (restricting comparison to the middle six months of the twelve-month period).

Figure 4 shows the proportions of false positive and false negative events, by broad cancer type (excluding non-melanoma skin cancer), measured in the cas2302 snapshot (the tumour groups are defined in Appendix 3). A more detailed tabulation is available by tumour group and tumour site in Appendix 5.

In most tumour groups, there are more tumours missed by the rapid registrations process (false negatives) than there are falsely identified as tumours (false positives).

For breast and prostate, very few incorrect proxy registrations are made. Breast, colorectal, lung, oesophagogastric (O-G) and prostate cancers are also least likely to be missing from the proxy dataset, whereas for cancers of unknown primary, and bone and soft tissue tumours more than 25% of cancers are missed. Bone and soft tissue tumours are not frequently diagnosed. These tumours often require multiple pathology reports to correctly diagnose a patient and the Rapid Registrations dataset has not attempted to reconcile differences in the reported diagnoses.

Figure 4: Types of error by tumour group

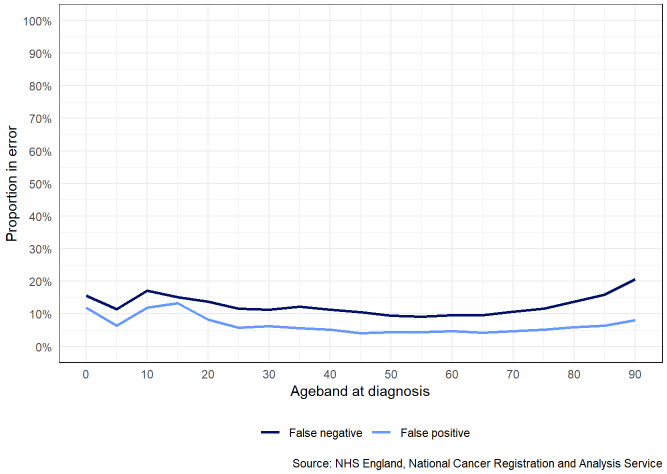


CNS: Central Nervous System; GI: Gastrointestinal; O-G: Oesophagogastric; ST: Soft Tissue
Source: NHS England, National Cancer Registration and Analysis Service

The proportion of false positive errors is fairly stable across all ages (Figure 5); the proportion of false negative errors slowly declines until age 70 when it increases significantly. The age dependence was investigated and the age-dependence of the basis of diagnosis was found to be at least partially responsible for this - see Appendix 6 for details.

The proportion of false positive cases is less sensitive to the age of the patient.

Figure 5: False negative and false positive errors by age band at diagnosis



Source: NHS England, National Cancer Registration and Analysis Service

The charts in Figure 6 (below) examine these patterns by tumour group. Please note that age groups for each tumour group must have a denominator of 25 patients or more or they are suppressed for reasons of statistical power.

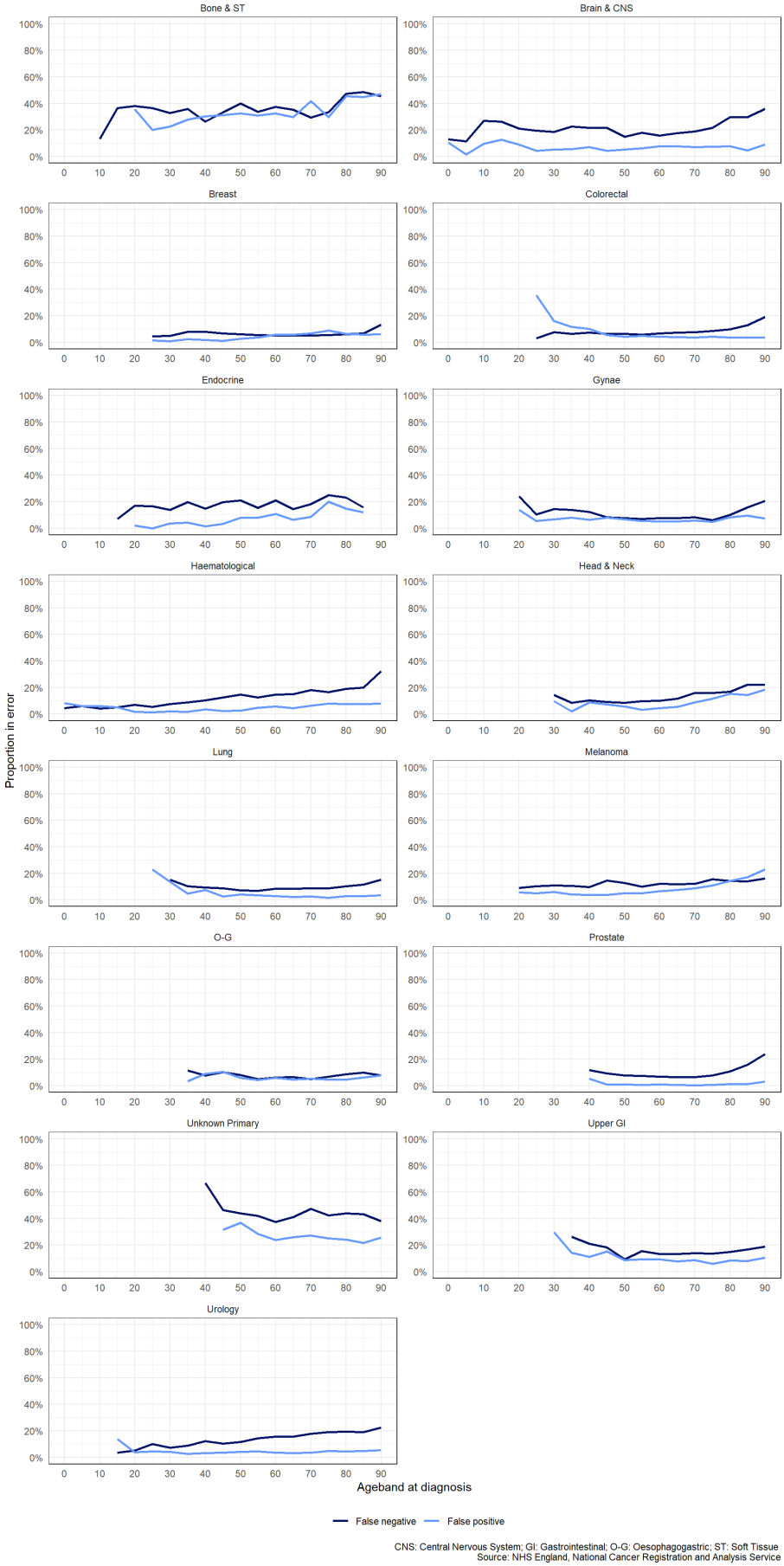
The patterns of false negative and false positive vary significantly by tumour group. Most groups have a higher proportion of false negatives than false positives at each age.

The proportion of false positives does not exhibit a trend by age for most tumour groups; the proportion rises with increasing age in the bone and soft tissue, head and neck groups and melanoma group and conversely falls with increasing age in the colorectal and unknown groups.

The proportion of false negatives rises with increasing age for all tumour groups except bone and soft tissue and endocrine. The most pronounced increases occur in the brain and central nervous system, colorectal, gynaecological, haematological, prostate, upper gastro-intestinal and unknown primary tumour groups.

The levels of both types of error are highest in tumour groups which are less likely to have solid-tissue pathology (haematological) or where survival rates are typically low. Conversely, the levels of error are lowest for tumour groups for which survival rates are typically higher.

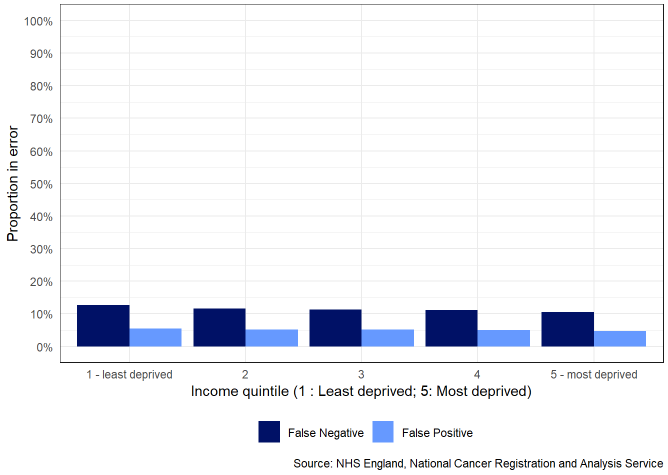
Figure 6: False negative and false positive errors by age band at diagnosis and tumour group



CNS: Central Nervous System; GI: Gastrointestinal; O-G: Oesophagogastric; ST: Soft Tissue
Source: NHS England, National Cancer Registration and Analysis Service

The variation of the false positive and false negative errors with Income deprivation quintile is shown in figure 7. While there is an overall trend visible this is likely to be due to confounding due to the variation with tumour type shown above and the known association of the incidence of many cancer types with income deprivation.

Figure 7: False negative and false positive errors by income deprivation quintile

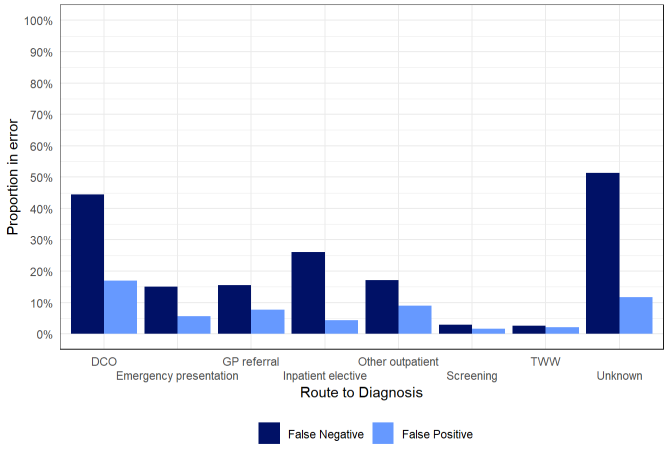


Source: NHS England, National Cancer Registration and Analysis Service

Figure 8 shows the variation of false negative and false positive errors with route to diagnosis. For false positives there is moderate variation with the lowest error rate being those cases identified through cancer screening or a two week wait referral. (These tumours are those that are likely to be captured in both the COSD dataset and the screening/Cancer Waiting Times datasets so the lower error rate is understandable.)

Most routes to diagnosis have a substantially higher false negative rate than the overall average. 'Two Week Wait' (TWW) and screening routes have a substantially lower false negative rate (and make up between them 45% of the total cohort).

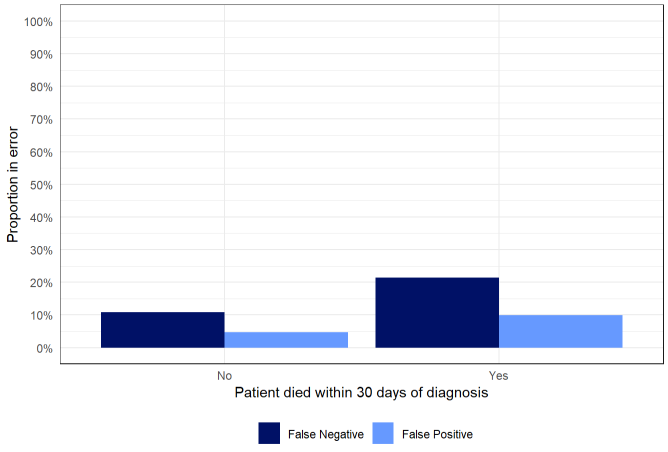
Figure 8: False negative and false positive errors by route to diagnosis



Source: NHS England, National Cancer Registration and Analysis Service

Figure 9 below shows the variation of false negative and false positive errors with whether or not the patient died within 30 days of diagnosis. The false negative error rate varies substantially between patients who die in the 30 days post-diagnosis compared to those who did, meaning that patients who die within 30 days are more likely to be missing from the dataset.

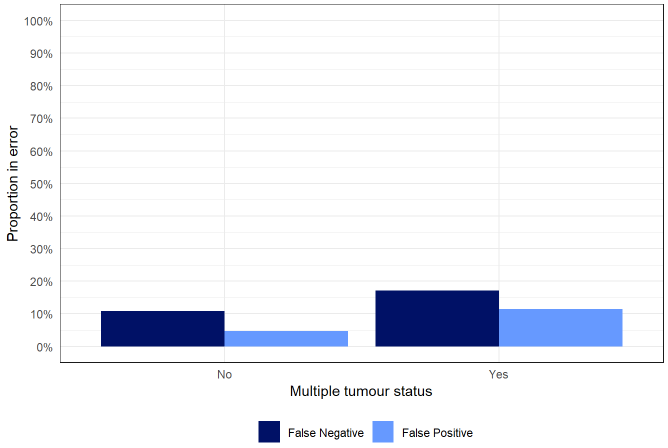
Figure 9: False negative and false positive errors by 30-day mortality



Source: NHS England, National Cancer Registration and Analysis Service

Figure 10 below shows the variation of false negative and false positive errors with the multiple tumour status of the patient, i.e. whether or not the patient had been diagnosed with more than one type of tumour in the period January 2018 onward. The false positive error rate varies substantially between patients with multiple tumour types and those that don't, meaning that these patients with multiple tumours are more likely to have incorrect tumour types or diagnosis dates recorded.

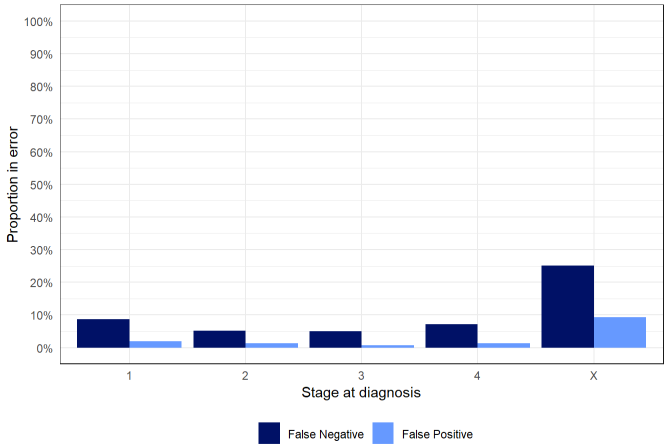
Figure 10: False negative and false positive errors by multiple tumour status



Source: NHS England, National Cancer Registration and Analysis Service

Figure 10b below shows the variation of false negative and false positive errors with the stage at diagnosis.

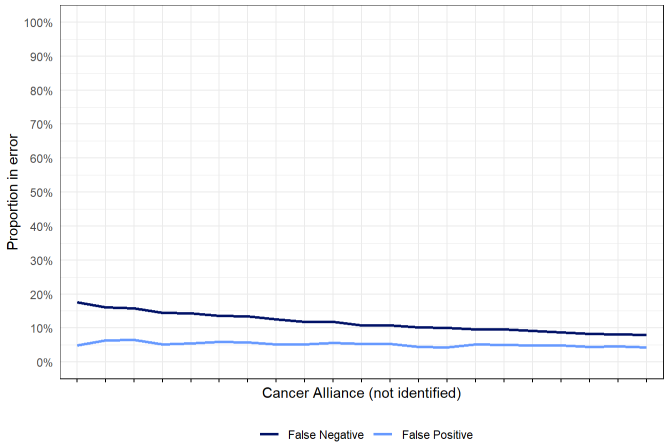
Figure 10b: False negative and false positive errors by stage



Source: NHS England, National Cancer Registration and Analysis Service

Figure 11 below shows the variation of false negative and false positive errors with the cancer alliance of residence of the patient at the time of diagnosis. The false negative error rate varies more in absolute terms than the false positive rate and may be driven by trust level variation (see figures 11 and 12 below).

Figure 11: False negative and false positive errors by Cancer Alliance



Source: NHS England, National Cancer Registration and Analysis Service

Figures 12 and 13 below show the variation of false negative and false positive errors with the trust that diagnosed the tumour. Figure 12 shows the error proportion and figure 13 the numerator (count) of the errors. Trusts shown are limited to NHS secondary care trusts with a denominator of at least 50 patients over the assessment period. Both figures are ordered in descending order of the false negative statistic - but note that the order is not the same in each figure.

There is substantial variation in both false positive and false negative rates and counts. Some large trusts have several hundred or up to 1000 cases (over the six-month period under assessment).

Figure 12: False negative and false positive errors (proportion) by hospital trust

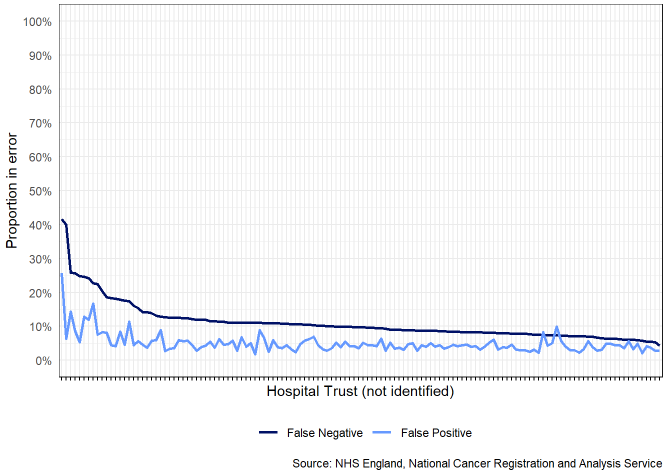
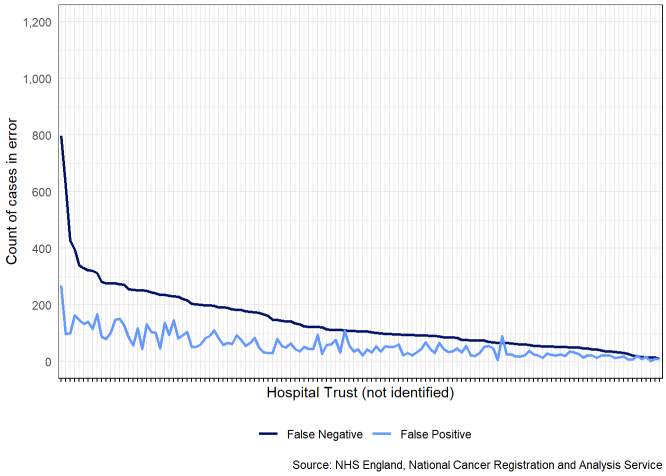


Figure 13: False negative and false positive errors (count) by hospital trust



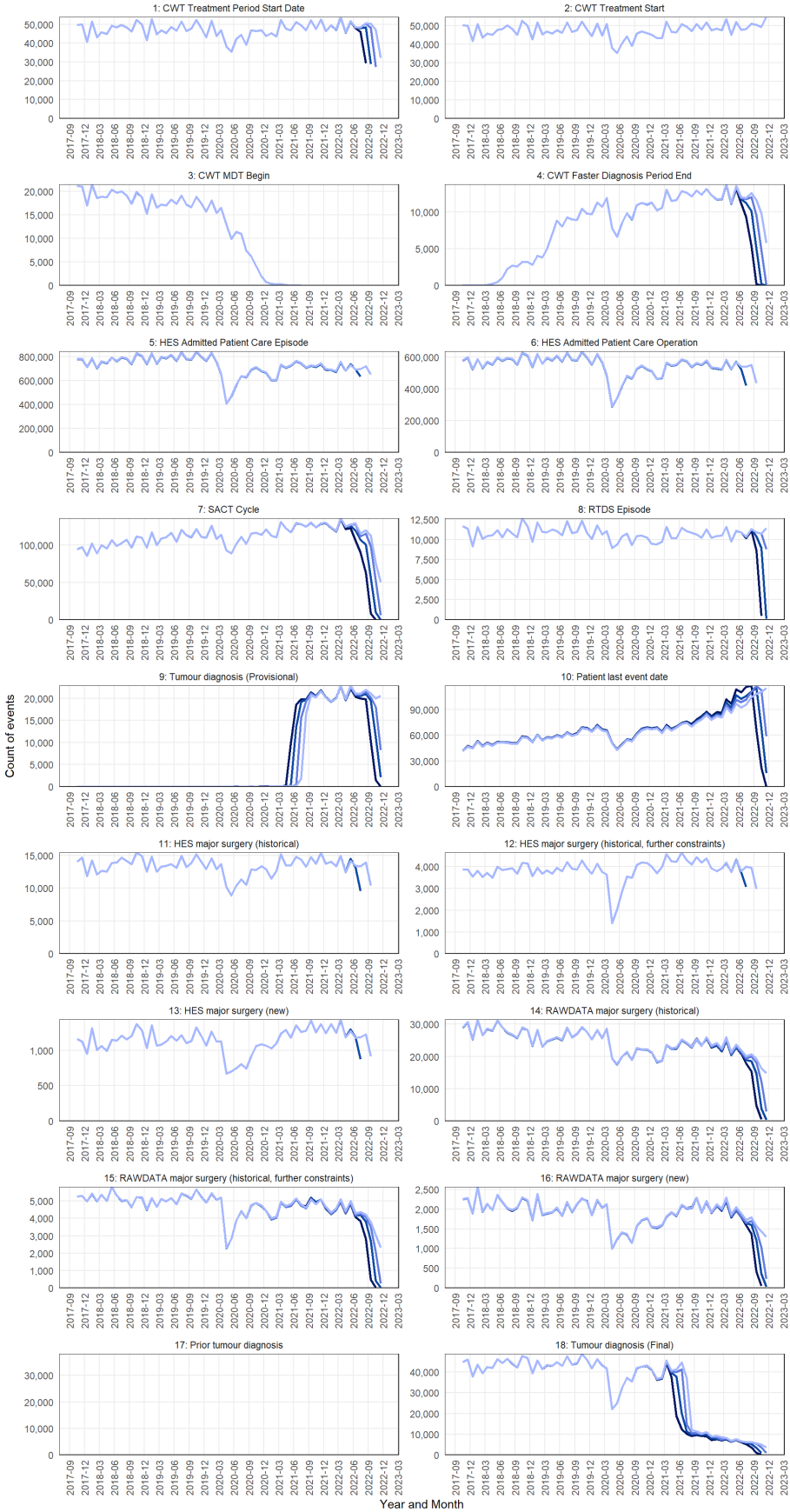
Counts of events over time

This section examines the population of events by chronological time and when they appear in successive analytical snapshots in the CAS. Figure 14 shows that most data items in the Rapid Registrations dataset are stable with respect to the snapshot month.

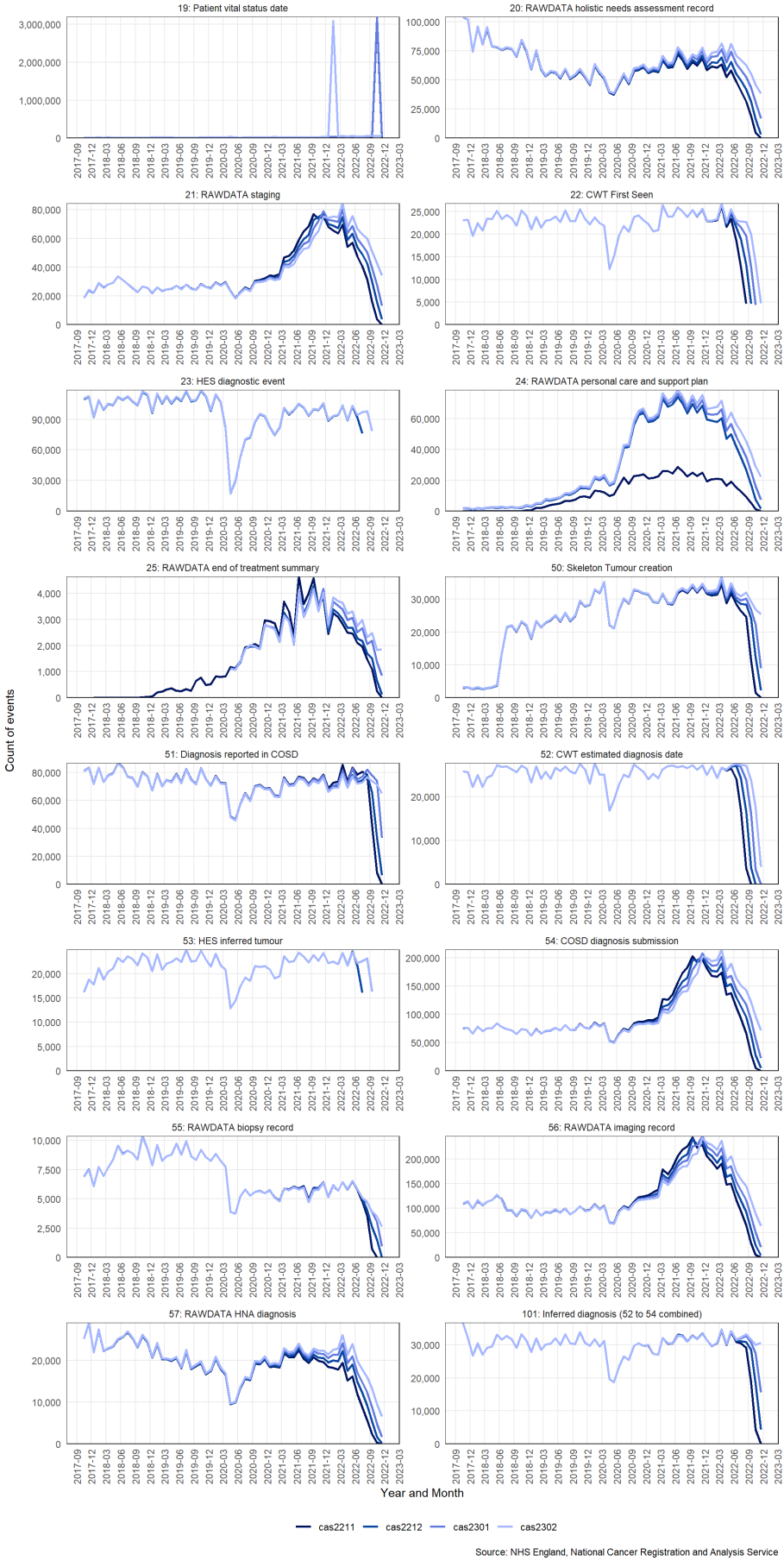
Specific comments about the events shown below are:

- Cancer Waiting Times data (events 1-4) are received based on the treatment start date, this explains the fact that for event 2 all lines lie exactly on top of each other. Other CWT events accumulate over successive snapshots where these events precede the first treatment start event.
- An issue with HES data resulting in lower than expected completeness port 2020-04-01 was resolved in cas2102, showing as increased event counts in events 5,6, 11, 12, 13 and 23.
- The definition of event 17 only includes tumour diagnoses prior to 2018, lack of data in the chart below is expected.
- Definitions of staging events may change between snapshots, this might explain higher or lower counts in one snapshot compared to others.
- The vital status shown in the event 19 is typically only assessed each January or the completion of registering each diagnosis year, explaining the large peaks in the graph.
- The raw data used to populate events 21, 54, and 56 is subject to ongoing deduplication, this explains lower counts in earlier time periods for later snapshots.
- Between snapshots there is generally an increase in the Event 101-103 (Inferred diagnoses) counts, particularly for recent months as additional COSD data is submitted. However, for some earlier months there is a small decrease in these event counts. This is because the algorithm to define Events 101-103 excludes potential diagnoses where the patient has a confirmed diagnosis for the same tumour group which was more than 90 days before the potential diagnosis, to avoid double-counting the same diagnosis. These exclusions can change between snapshots due to the processing of gold standard cancer registration data, which leads to an increase in confirmed previous diagnoses. However the magnitude of this effect has been measured to be <1% of all cases in any given month.

Figure 14: Population of data items to CAS snapshot



Source: NHS England, National Cancer Registration and Analysis Service



Source: NHS England, National Cancer Registration and Analysis Service

Estimated completeness of Rapid Registrations and secondary datasets

Detailed linked rapid cancer registration, CWT, SACT and RTDS data is available at approximately a four-month lag from real time. Linked HES and raw COSD data is available at approximately 4-5 months behind real time.

Table 2 below shows data usability and completeness for Rapid Registrations and the constituent datasets. The “latest usable” column shows the ‘hard limit’ on data that is considered fit for analytical purposes (90% completeness), even in months prior to this though data is not necessarily considered complete and the completeness is displayed below. This should be taken into account in any use of the rapid registration data and the secondary datasets.

For the Rapid Tumour data completeness is expressed as the proportion of CCG of residence which show a cancer incidence within the normally expected range (see Table 3 below). For other datasets except CWT completeness is computed as a percentage of the number of data providers who have supplied data over those who are expected to do so.

Data completeness within the Cancer Waiting Times dataset varies at patient level with event type. Figures for the Treatment Start Date and Treatment Period Start Date are given below. Completeness of other CWT events can be estimated by inspecting Figure 13 (events 1-4).

Table 2: Rapid registration and dataset usability/completeness in cas2302

Data source	Latest usable	March 2022	April 2022	May 2022	June 2022	July 2022	August 2022	September 2022	October 2022	November 2022
Rapid Tumours (COSD)	November 2022	Complete	97%	Complete	Complete	97%	97%	93%	92%	94%
HES	August 2022	Complete	Complete	Complete	Complete	Complete	Complete	•	•	•
SACT	July 2022	98%	97%	95%	96%	92%	•	•	•	•
RTDS	September 2022	96%	94%	92%	96%	94%	96%	94%	•	•
CWT (TSD)	November 2022	Complete	Complete	Complete	Complete	Complete	Complete	Complete	Complete	Complete
CWT (TPSD)	October 2022	Complete	Complete	Complete	Complete	Complete	Complete	Complete	98%	•

Note:
COSD = Cancer Outcomes and Services Dataset
TSD = Treatment Start Date
TPSD = Treatment Period Start Date

Table 3: Number of outlier CCGs in COSD dataset in cas2302

The table below shows the number of CCGs (using the April 2020 boundaries) which have 3-sigma outlier counts per month (either high or low) compared to the expectation of the fraction of the total number of new cancer registrations in England. This can be used to judge to what extent there is large scale missing data in COSD (and therefore in the Rapid Registrations in any particular month.)

Year and month	Outlier: High	Outlier: Low	In expected range	Total received
2020-01	0	1	134	135
2020-02	1	0	134	135
2020-03	0	1	134	135
2020-04	4	7	124	135
2020-05	4	2	129	135
2020-06	1	3	131	135
2020-07	1	0	134	135
2020-08	1	4	130	135
2020-09	1	0	134	135
2020-10	0	4	131	135
2020-11	0	1	134	135
2020-12	1	1	133	135
2021-01	0	0	135	135
2021-02	1	2	132	135
2021-03	2	2	131	135
2021-04	2	0	133	135
2021-05	0	1	134	135
2021-06	0	1	134	135
2021-07	0	1	134	135
2021-08	0	1	134	135
2021-09	2	3	130	135
2021-10	1	2	132	135
2021-11	0	1	134	135
2021-12	0	1	134	135
2022-01	2	4	129	135
2022-02	0	2	133	135
2022-03	0	3	132	135
2022-04	0	5	130	135

Year and month	Outlier: High	Outlier: Low	In expected range	Total received
2022-05	1	1	133	135
2022-06	1	1	133	135
2022-07	1	2	132	135
2022-08	1	3	131	135
2022-09	1	9	125	135
2022-10	5	6	124	135
2022-11	0	8	127	135
2022-12	35	38	52	125

Staging data in the Rapid Registrations dataset

TNM stage group 1-4

The size and extent of a cancer is commonly described using the 'TNM' system (https://www.uicc.org/resources/tnm) for "Tumour", "Node", and "Metastases". This is often abbreviated to a number between 1 (typically a localised tumour with limited spread) to 4 (typically a tumour that has invaded or spread to distant organs). The stage at diagnosis is very strongly associated with patient outcomes.

In the current version of the Rapid Registrations dataset partial staging data is provided for a number of different cancer sites (ICD-10 codes can be found in the labels for tables 5a-k). This has been benchmarked against the gold standard cancer registry data for cas2302.

Table 4 shows the count and proportion of cases by TNM stage group for both the Rapid Registrations and the Gold Standard Registrations, for calendar year 2018. For example 32% of breast cancers are TNM stage group 1 in the Rapid Registrations, but 38% in the Gold Standard Registrations. Compared to the Gold Standard Registrations in 2018, the Rapid Registrations under report breast cancers diagnosed at stages 1 or 2; colorectal cancers diagnosed at stage 4 are under reported and prostate cancers have under reported stages 1 and 4. In all three tumour groups, there are more tumours allocated to the unknown or unstageable category. Lung cancers in the RCRD most accurately match the Gold Standard Registrations and exhibits a broadly similar stage profile from both measures.

Table 4: Summary proportions of stage at diagnosis for the Rapid Registrations and Gold Standard Registrations

Broad Cancer Group	Stage Group	Count (Rapid)	Percentage (Rapid)	Count (Gold Standard)	Percentage (Gold Standard)
Bladder	1	2321	24.2%	2869	29.9%
Bladder	2	1799	18.7%	1879	19.6%
Bladder	3	559	5.8%	885	9.2%
Bladder	4	258	2.7%	659	6.9%
Bladder	U	4665	48.6%	3310	34.5%
Breast	1	14049	31.8%	16576	37.5%
Breast	2	13247	30.0%	16734	37.9%
Breast	3	3235	7.3%	3689	8.4%
Breast	4	1185	2.7%	1974	4.5%
Breast	U	12463	28.2%	5206	11.8%
Colorectum	1	4918	15.0%	5506	16.8%
Colorectum	2	7037	21.4%	7725	23.5%
Colorectum	3	8244	25.1%	9311	28.4%
Colorectum	4	5116	15.6%	7477	22.8%
Colorectum	U	7526	22.9%	2822	8.6%
Kidney	1	2383	28.8%	3348	40.5%
Kidney	2	447	5.4%	558	6.8%
Kidney	3	1370	16.6%	1660	20.1%
Kidney	4	686	8.3%	1581	19.1%
Kidney	U	3374	40.8%	1113	13.5%
Lung	1	6176	17.1%	6647	18.4%
Lung	2	2587	7.2%	2694	7.5%
Lung	3	7306	20.2%	7618	21.1%
Lung	4	14925	41.3%	17213	47.7%
Lung	U	5120	14.2%	1942	5.4%
Lymphoma	1	909	7.4%	1756	14.4%
Lymphoma	2	951	7.8%	1623	13.3%
Lymphoma	3	1201	9.8%	2002	16.4%
Lymphoma	4	2658	21.7%	4946	40.4%
Lymphoma	U	6512	53.2%	1904	15.6%
Melanoma	1	6335	48.0%	8264	62.7%

Broad Cancer Group	Stage Group	Count (Rapid)	Percentage (Rapid)	Count (Gold Standard)	Percentage (Gold Standard)
Melanoma	2	2389	18.1%	2653	20.1%
Melanoma	3	444	3.4%	1034	7.8%
Melanoma	4	203	1.5%	350	2.7%
Melanoma	U	3817	28.9%	887	6.7%
Oesophagus	1	291	3.5%	449	5.4%
Oesophagus	2	1506	18.0%	971	11.6%
Oesophagus	3	1785	21.4%	2156	25.8%
Oesophagus	4	2554	30.6%	3251	38.9%
Oesophagus	U	2211	26.5%	1520	18.2%
Ovary	1	1151	22.5%	1480	29.0%
Ovary	2	234	4.6%	279	5.5%
Ovary	3	1182	23.1%	1632	31.9%
Ovary	4	692	13.5%	1051	20.6%
Ovary	U	1849	36.2%	666	13.0%
Pancreas	1	361	4.5%	670	8.3%
Pancreas	2	618	7.7%	804	10.0%
Pancreas	3	750	9.3%	1039	12.9%
Pancreas	4	2040	25.4%	4125	51.4%
Pancreas	U	4255	53.0%	1386	17.3%
Prostate	1	11630	25.1%	16273	35.1%
Prostate	2	5542	11.9%	6571	14.2%
Prostate	3	10403	22.4%	11685	25.2%
Prostate	4	5640	12.2%	8105	17.5%
Prostate	U	13194	28.4%	3775	8.1%
Stomach	1	317	8.3%	334	8.7%
Stomach	2	358	9.3%	452	11.8%
Stomach	3	608	15.8%	679	17.7%
Stomach	4	1100	28.7%	1620	42.2%
Stomach	U	1455	37.9%	753	19.6%
Uterus	1	4645	58.1%	5417	67.7%
Uterus	2	513	6.4%	544	6.8%
Uterus	3	732	9.1%	823	10.3%
Uterus	4	505	6.3%	559	7.0%
Uterus	U	1606	20.1%	658	8.2%

In Tables 5a-m below, the distribution of the stage allocations between the Rapid Registrations and the Gold Standard Registrations are examined.

The figures indicate the proportion of agreement at the 1-digit TNM stage group level, where the stage is known in the Rapid Registrations dataset. Stages 1-4 in the Rapid Registrations dataset agree with the gold standard stage variable for a high proportion.

For example, when examining the subset of Rapid Registrations breast tumours that are identified as TNM stage 1 (32%), approximately 89% of these are found to be TNM stage group 1 in the gold standard registration data, with another 11% distributed across TNM stages 2-4 and the unknown or unstageable groups.

For many but not all (e.g., late stage breast cancer), roughly 85% or more of staged cases in the Rapid Registrations table have the same stage grouping as the equivalent tumour in the standard registration data - this can be seen in the table below by inspecting the figures where the stage metrics for the Rapid Registrations and Gold Standard Registrations are the same.

Where the stage is labelled as unknown or unstageable in the rapid pathway dataset it is known for at least 70% of those cases in the gold standard data.

Tables 5a-m: Stage comparison between Rapid Registrations and Gold Standard Registrations by cancer site

a. bladder (ICD-10 C67)

	Stage Group (Rapid)				
Stage Group (Gold Standard)	**1**	**2**	**3**	**4**	**Unknown**
1	84.8%	4.2%	7.9%	5.4%	16.4%
2	3.8%	71.7%	15.7%	5.8%	8.5%
3	2.6%	10.9%	64.9%	4.7%	5.4%
4	1.2%	4.9%	5.5%	79.1%	6.6%
U	7.5%	8.3%	5.9%	5.0%	63.0%

b. breast (ICD-10 C50)

	Stage Group (Rapid)				
Stage Group (Gold Standard)	**1**	**2**	**3**	**4**	**Unknown**
1	89.1%	4.9%	1.5%	3.4%	26.7%
2	6.5%	88.5%	10.9%	14.3%	28.6%
3	0.5%	2.7%	80.3%	5.5%	4.8%
4	0.2%	0.9%	2.9%	71.8%	7.1%
U	3.7%	3.0%	4.3%	5.0%	32.8%

c. colorectum (ICD-10 C18-C20)

	Stage Group (Rapid)				
Stage Group (Gold Standard)	**1**	**2**	**3**	**4**	**Unknown**
1	84.9%	2.1%	1.8%	0.7%	13.3%
2	5.7%	85.6%	5.5%	1.2%	12.0%
3	6.6%	7.5%	85.1%	4.4%	16.2%
4	0.9%	2.8%	5.8%	92.7%	26.7%
U	1.9%	2.0%	1.8%	1.0%	31.8%

d. kidney (ICD-10 C64)

	Stage Group (Rapid)				
Stage Group (Gold Standard)	**1**	**2**	**3**	**4**	**Unknown**
1	91.2%	6.7%	3.1%	1.7%	32.3%
2	0.5%	78.3%	1.0%	0.7%	5.2%
3	1.8%	6.7%	85.7%	3.9%	11.5%
4	0.5%	3.4%	6.0%	92.4%	24.9%
U	6.1%	4.9%	4.2%	1.2%	26.1%

e. lung (ICD-10 C33-C34)

	Stage Group (Rapid)				
Stage Group (Gold Standard)	**1**	**2**	**3**	**4**	**Unknown**
1	93.7%	6.5%	1.1%	0.4%	10.6%
2	2.7%	84.6%	1.8%	0.3%	3.1%
3	1.7%	4.8%	90.7%	1.3%	11.2%
4	1.2%	3.1%	5.5%	97.5%	41.2%
U	0.8%	1.0%	0.9%	0.5%	33.8%

f. melanoma (ICD-10 C43)

	Stage Group (Rapid)				
Stage Group (Gold Standard)	**1**	**2**	**3**	**4**	**Unknown**
1	94.3%	1.8%	5.9%	8.9%	57.8%
2	2.1%	79.0%	9.0%	18.2%	14.5%
3	1.9%	11.7%	78.2%	15.3%	6.6%
4	0.2%	1.6%	2.5%	46.3%	5.2%
U	1.5%	5.9%	4.5%	11.3%	15.9%

g. oesophagus (ICD-10 C15)

	Stage Group (Rapid)				
Stage Group (Gold Standard)	**1**	**2**	**3**	**4**	**Unknown**
1	80.8%	5.0%	0.5%	0.2%	5.6%
2	7.9%	49.5%	3.5%	1.0%	5.1%
3	2.1%	35.0%	68.6%	6.3%	10.7%

	Stage Group (Rapid)				
Stage Group (Gold Standard)	**1**	**2**	**3**	**4**	**Unknown**
4	1.0%	5.4%	21.8%	83.4%	29.3%
U	8.2%	5.0%	5.5%	9.1%	49.3%

h. ovary (ICD-10 C56-C57)

	Stage Group (Rapid)				
Stage Group (Gold Standard)	**1**	**2**	**3**	**4**	**Unknown**
1	97.3%	7.3%	0.9%	0.3%	17.8%
2	0.4%	88.0%	0.5%	NA	3.4%
3	0.8%	2.6%	91.5%	11.1%	24.8%
4	0.3%	0.4%	4.4%	84.4%	22.2%
U	1.2%	1.7%	2.6%	4.2%	31.8%

i. prostate (ICD-10 C61)

	Stage Group (Rapid)				
Stage Group (Gold Standard)	**1**	**2**	**3**	**4**	**Unknown**
1	86.3%	9.6%	4.2%	1.3%	39.3%
2	6.7%	82.9%	2.5%	0.9%	6.7%
3	4.3%	4.2%	86.6%	2.7%	13.6%
4	0.8%	0.8%	4.0%	93.1%	17.5%
U	2.0%	2.5%	2.6%	2.0%	22.9%

j. stomach (ICD-10 C16)

	Stage Group (Rapid)				
Stage Group (Gold Standard)	**1**	**2**	**3**	**4**	**Unknown**
1	67.5%	4.7%	0.7%	0.1%	6.7%
2	19.2%	66.5%	10.2%	0.8%	5.6%
3	6.0%	18.2%	69.7%	3.1%	9.4%
4	1.9%	6.4%	15.5%	94.0%	31.8%
U	5.4%	4.2%	3.9%	2.0%	46.4%

k. uterus (ICD-10 C54-C55)

	Stage Group (Rapid)				
Stage Group (Gold Standard)	**1**	**2**	**3**	**4**	**Unknown**
1	97.6%	10.9%	5.7%	7.3%	46.6%
2	0.6%	83.6%	1.2%	2.2%	4.2%
3	0.5%	2.1%	87.8%	6.5%	7.0%
4	0.2%	1.8%	2.3%	77.2%	8.3%
U	1.1%	1.6%	2.9%	6.7%	33.9%

l. pancreas (ICD-10 C25)

	Stage Group (Rapid)				
Stage Group (Gold Standard)	**1**	**2**	**3**	**4**	**Unknown**
1	73.7%	3.6%	0.9%	0.3%	8.6%
2	14.7%	75.2%	2.4%	0.5%	6.0%
3	4.7%	12.0%	88.7%	0.6%	6.4%
4	3.3%	6.0%	6.1%	97.6%	47.9%
U	3.6%	3.2%	1.9%	0.9%	31.0%

m. lymphoma (ICD-10 C81-C86, C88)

	Stage Group (Rapid)				
Stage Group (Gold Standard)	**1**	**2**	**3**	**4**	**Unknown**

	Stage Group (Rapid)				
Stage Group (Gold Standard)	**1**	**2**	**3**	**4**	**Unknown**
1	90.4%	1.3%	0.5%	0.5%	13.9%
2	0.9%	93.3%	1.2%	0.5%	10.7%
3	0.4%	1.3%	90.3%	1.5%	13.3%
4	5.8%	2.6%	7.0%	93.1%	35.5%
U	2.4%	1.6%	1.0%	4.4%	26.7%

“Early” vs “Late” stage

Below in table 6 we repeat the above tabulations but now grouping Rapid and Gold Standard cancers into “Early” (TNM stage group 1 & 2) or “Late” (TNM stage group 3 & 4) categories. We see that 62% of breast cancers are identified as “Early” stage in the Rapid Registrations dataset compared to 76% in the Gold Standard Registration data due to the higher proportion of “Unknown” stage tumours (28% vs 10% respectively).

As with the more detailed stage data, there is a high degree of concordance between the gold standard and rapid registration stage fields if a known stage can be identified.

Table 6: Summary proportions of “Early” vs “Late” stage for Rapid Registrations and Gold Standard Registrations

Broad Cancer Group	Stage Group	Count (Rapid)	Percentage (Rapid)	Count (Gold Standard)	Percentage (Gold Standard)
Bladder	Early	4120	42.9%	4748	49.4%
Bladder	Late	817	8.5%	1544	16.1%
Bladder	Unknown	4665	48.6%	3310	34.5%
Breast	Early	27296	61.8%	33310	75.4%
Breast	Late	4420	10.0%	5663	12.8%
Breast	Unknown	12463	28.2%	5206	11.8%
Colorectum	Early	11955	36.4%	13231	40.3%
Colorectum	Late	13360	40.7%	16788	51.1%
Colorectum	Unknown	7526	22.9%	2822	8.6%
Kidney	Early	2830	34.3%	3906	47.3%
Kidney	Late	2056	24.9%	3241	39.2%
Kidney	Unknown	3374	40.8%	1113	13.5%
Lung	Early	8763	24.3%	9341	25.9%
Lung	Late	22231	61.6%	24831	68.8%
Lung	Unknown	5120	14.2%	1942	5.4%
Lymphoma	Early	1860	15.2%	3379	27.6%
Lymphoma	Late	3859	31.6%	6948	56.8%
Lymphoma	Unknown	6512	53.2%	1904	15.6%
Melanoma	Early	8724	66.2%	10917	82.8%
Melanoma	Late	647	4.9%	1384	10.5%
Melanoma	Unknown	3817	28.9%	887	6.7%
Oesophagus	Early	1797	21.5%	1420	17.0%
Oesophagus	Late	4339	52.0%	5407	64.8%
Oesophagus	Unknown	2211	26.5%	1520	18.2%
Ovary	Early	1385	27.1%	1759	34.4%
Ovary	Late	1874	36.7%	2683	52.5%
Ovary	Unknown	1849	36.2%	666	13.0%
Pancreas	Early	979	12.2%	1474	18.4%
Pancreas	Late	2790	34.8%	5164	64.4%
Pancreas	Unknown	4255	53.0%	1386	17.3%
Prostate	Early	17172	37.0%	22844	49.2%
Prostate	Late	16043	34.6%	19790	42.6%
Prostate	Unknown	13194	28.4%	3775	8.1%
Stomach	Early	675	17.6%	786	20.5%
Stomach	Late	1708	44.5%	2299	59.9%
Stomach	Unknown	1455	37.9%	753	19.6%
Uterus	Early	5158	64.5%	5961	74.5%
Uterus	Late	1237	15.5%	1382	17.3%

Broad Cancer Group	Stage Group	Count (Rapid)	Percentage (Rapid)	Count (Gold Standard)	Percentage (Gold Standard)
Uterus	Unknown	1606	20.1%	658	8.2%

In Table 7a-m below the distribution of the stage allocation between the Rapid Registrations and the Gold Standard Registrations are examined, aggregated into Early and Late stage.

Tables 7a-m: “Early” vs “late” stage comparison between Rapid Registrations and Gold Standard Registrations

a. bladder (ICD-10 C67)

	Stage Category (Rapid)		
Stage Category (Gold Standard)	**Early**	**Late**	**Unknown**
Early	83.1%	19.7%	25.0%
Late	9.1%	74.7%	12.0%
Unknown	7.8%	5.6%	63.0%

b. breast (ICD-10 C50)

	Stage Category (Rapid)		
Stage Category (Gold Standard)	**Early**	**Late**	**Unknown**
Early	94.5%	13.9%	55.3%
Late	2.1%	81.6%	11.9%
Unknown	3.4%	4.5%	32.8%

c. colorectum (ICD-10 C18-C20)

	Stage Category (Rapid)		
Stage Category (Gold Standard)	**Early**	**Late**	**Unknown**
Early	88.9%	5.2%	25.3%
Late	9.2%	93.3%	42.9%
Unknown	1.9%	1.5%	31.8%

d. kidney (ICD-10 C64)

	Stage Category (Rapid)		
Stage Category (Gold Standard)	**Early**	**Late**	**Unknown**
Early	90.6%	3.6%	37.6%
Late	3.5%	93.2%	36.3%
Unknown	5.9%	3.2%	26.1%

e. lung (ICD-10 C33-C34)

	Stage Category (Rapid)		
Stage Category (Gold Standard)	**Early**	**Late**	**Unknown**
Early	94.8%	1.5%	13.8%
Late	4.4%	97.9%	52.4%
Unknown	0.8%	0.6%	33.8%

f. melanoma (ICD-10 C43)

	Stage Category (Rapid)		
Stage Category (Gold Standard)	**Early**	**Late**	**Unknown**
Early	92.1%	18.7%	72.3%
Late	5.2%	74.7%	11.8%
Unknown	2.7%	6.6%	15.9%

g. Oesophagus (ICD-10 C15)

	Stage Category (Rapid)		
Stage Category (Gold Standard)	**Early**	**Late**	**Unknown**
Early	60.1%	2.4%	10.7%
Late	34.3%	90.0%	40.0%
Unknown	5.6%	7.6%	49.3%

h. ovary (ICD-10 C56-C57)

	Stage Category (Rapid)		
Stage Category (Gold Standard)	**Early**	**Late**	**Unknown**
Early	97.3%	1.0%	21.2%
Late	1.4%	95.8%	47.0%
Unknown	1.3%	3.2%	31.8%

i. prostate (ICD-10 C61)

	Stage Category (Rapid)		
Stage Category (Gold Standard)	**Early**	**Late**	**Unknown**
Early	92.8%	5.2%	46.1%
Late	5.0%	92.4%	31.1%
Unknown	2.1%	2.4%	22.9%

j. stomach (ICD-10 C16)

	Stage Category (Rapid)		
Stage Category (Gold Standard)	**Early**	**Late**	**Unknown**
Early	78.5%	4.4%	12.4%
Late	16.7%	92.9%	41.2%
Unknown	4.7%	2.7%	46.4%

k. uterus (ICD-10 C54-C55)

	Stage Category (Rapid)		
Stage Category (Gold Standard)	**Early**	**Late**	**Unknown**
Early	97.8%	8.0%	50.8%
Late	1.0%	87.6%	15.3%
Unknown	1.1%	4.4%	33.9%

l. pancreas (ICD-10 C25)

	Stage Category (Rapid)		
Stage Category (Gold Standard)	**Early**	**Late**	**Unknown**
Early	82.3%	1.5%	14.7%
Late	14.3%	97.3%	54.3%
Unknown	3.4%	1.2%	31.0%

m. lymphoma (ICD-10 C81-C86, C88)

	Stage Category (Rapid)		
Stage Category (Gold Standard)	**Early**	**Late**	**Unknown**
Early	93.0%	1.3%	24.6%
Late	5.1%	95.4%	48.7%
Unknown	2.0%	3.3%	26.7%

Stage trends over time

Figure 15 shows the monthly variation of the incidence count by stage at diagnosis for a number of common cancers. Allowing for variation in the number of working days in each month (which affects the overall number of tumours diagnosed per month) and for statistical fluctuation there is little evidence of any stage shift in the period displayed. The feature around May 2018 in the prostate cancer trends can be ascribed to the so called 'Turnbull-Fry effect' (https://www.ndrs.nhs.uk/examining-the-fry-and-turnbull-effect-on-prostate-cancer-incidence-in-england/).

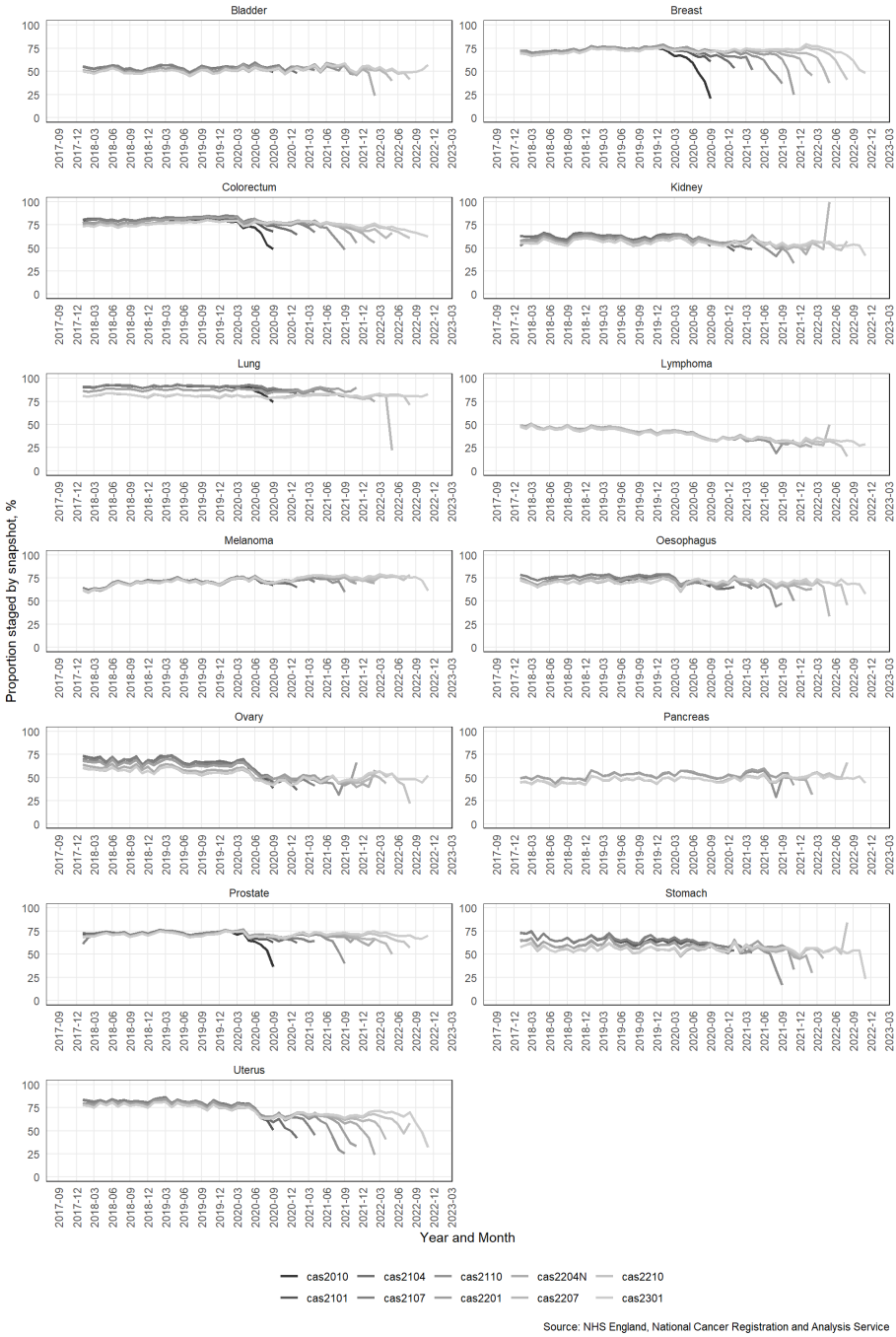
Figure 15: Stage trends over time

Source: NHS England, National Cancer Registration and Analysis Service

Stage completeness by snapshot

Figure 16 shows the completeness of stage by tumour type for one snapshot per quarter. Stage completeness continues to increase and lags behind the incidence completeness due to staging activity happening up to several months after diagnosis.

Figure 16: Stage completeness by snapshot

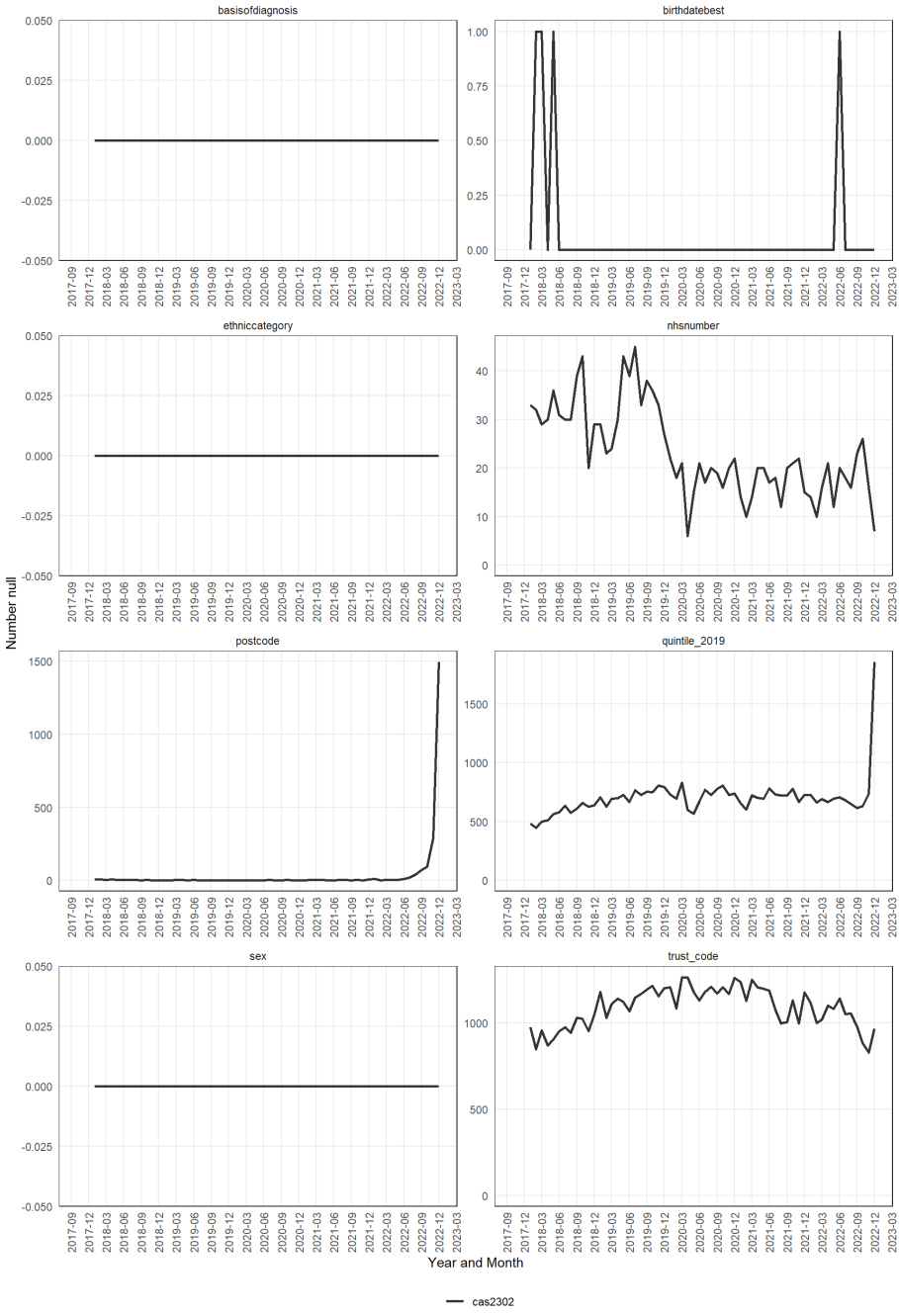


Source: NHS England, National Cancer Registration and Analysis Service

Counts of missing data

Figure 17 shows the count of tumours per month where the indicated data item is missing. The data items are: basis of diagnosis, birth date best, ethnic category, NHS number, postcode, quintile 2019, sex and trust code. Larger counts in the most recent months are to be expected.

Figure 17: Counts of missing data

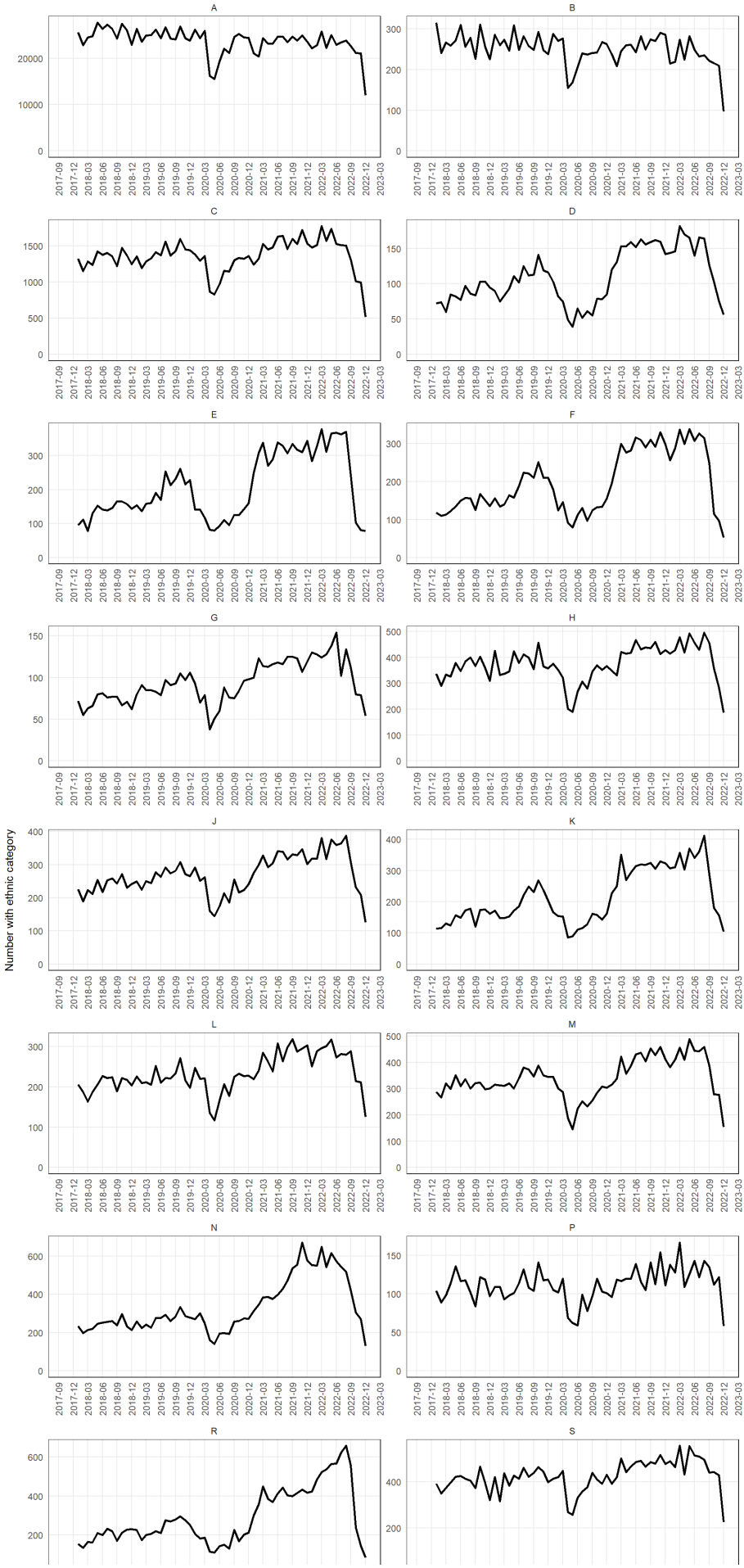


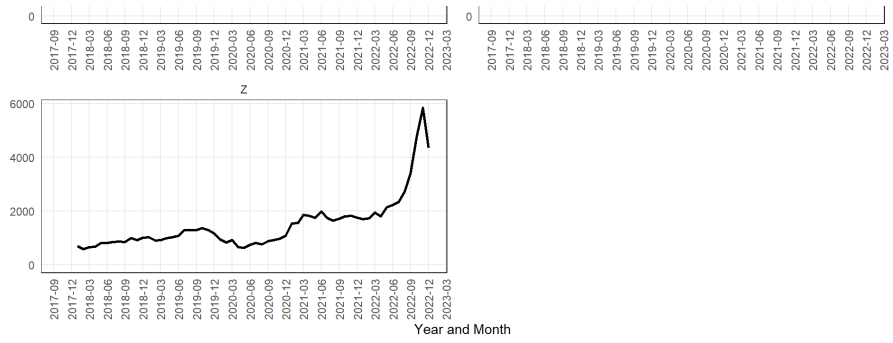
Source: NHS England, National Cancer Registration and Analysis Service

Ethnicity completeness

Figure 18 shows the count of tumours per month where the indicated data item is missing. Larger counts in the most recent months are to be expected.

Figure 18: Ethnicity completeness


A
B
C
D
E
F
G
H
J
K
L
M
N
P
R
S
Number with ethnic category

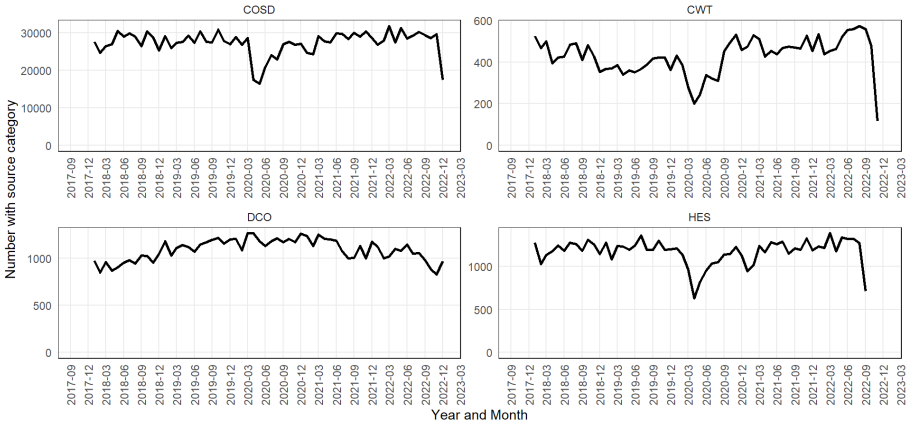


Source: NHS England, National Cancer Registration and Analysis Service

Tumour source

Figure 19 shows the proportion of tumours created by the source of the diagnosis - i.e., which dataset was used to create them, by month

Figure 19: Tumour source dataset

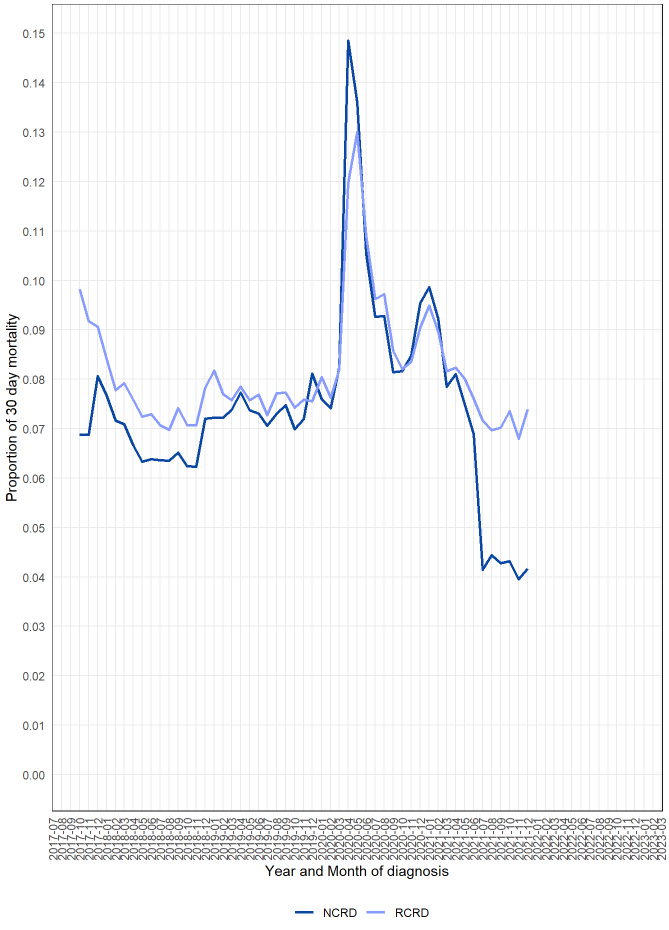


Source: NHS England, National Cancer Registration and Analysis Service

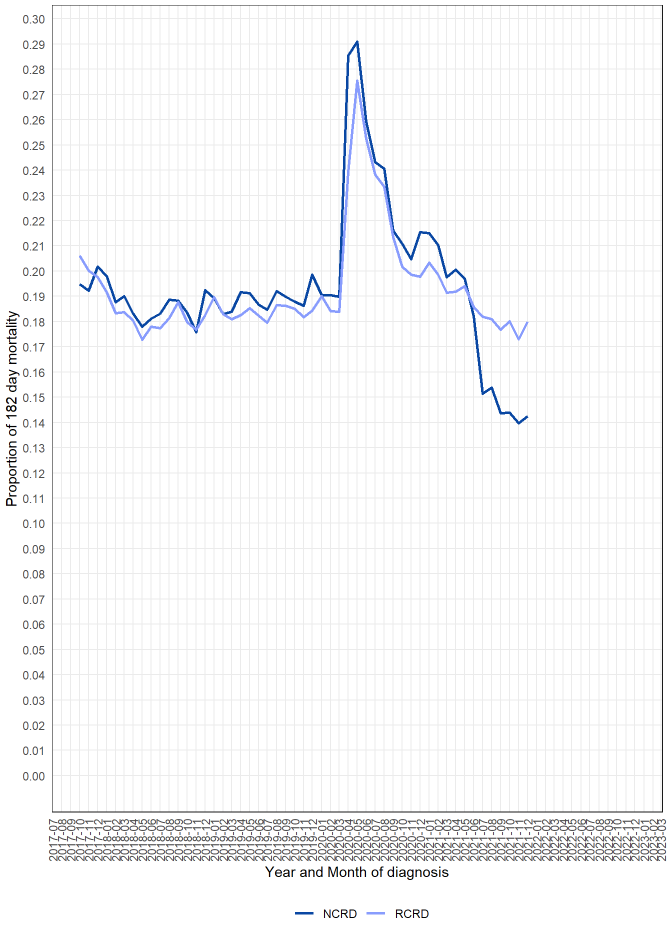
Mortality proportion by month

Figure 20 shows the mortality proportions by month mortality within 30 and 182 days in the RCRD compared to the NCRD, for all cancers included in RCRD excl C44 and D06.

Figure 20: Monthly mortality proportions at 30 and 182 days,



Source: NHS England, National Cancer Registration and Analysis Service



Source: NHS England, National Cancer Registration and Analysis Service

Appendix 1 - List of pathway events

Table A1: AT_RAPID_PATHWAY: event list

EVENT_TYPE	EVENT_DESC	EVENT_PROPERTY_1	EVENT_PROPERTY_2	EVENT_PROPERTY_3	EVENT_DATE	Linkage
1	CWT Treatment Period Start Date	CWT First Treatment Flag	CWT SITE_ICD10	CWT Cancer Treatment Event Type	Treat period start	NHSNUMBER
2	CWT Treatment Start	CWT Treatment Modality	CWT Cancer Treatment Event type		Treatment start date	NHSNUMBER
3	CWT MDT Begin	CWT MDT Cancer Care Plan discussed indicator			MDT date	NHSNUMBER
4	CWT Faster Diagnosis Period End	(null)	Faster Diagnosis Period site		Faster Diagnosis Period end date	NHSNUMBER
5	HES Admitted Patient Care Episode	Treatment speciality	All ICD-10 codes (for episode)	All OPCS-4 codes (for episode)	Episode Start date - Episode end date	NHSNUMBER
6	HES Admitted Patient Care Operation	OPCS codes (for date) in POS order	ICD-10 codes (for episode)		Operation date	NHSNUMBER
7	SACT Cycle	Benchmark group	Cycle number	Treatment intent	Cycle start date	PATIENTID
8	RTDS Episode	Radiotherapy intent	ICD-10 diagnosis code		Episode treatment start date	PATIENTID
9	Tumour diagnosis (Provisional)	Statusofregistration	ICD-10 diagnosis code	Stage_best	Diagnosisdatebest	PATIENTID
10	Patient last event date	Vitalstatus			Dateofvitalstatus1 (start of range)	PATIENTID
11	HES major surgery (historical)	OPCS-4 code	ICD-10 diagnosis code	Further notes/constraints	Operation date	NHSNUMBER
12	HES major surgery (historical, further constraints)	OPCS-4 code	ICD-10 diagnosis code	Further notes/constraints	Operation date	NHSNUMBER
13	HES major surgery (new)	OPCS-4 code	ICD-10 diagnosis code	Further notes/constraints	Operation date	NHSNUMBER
14	RAWDATA major surgery (historical)	OPCS-4 code	ICD-10 diagnosis code	Further notes/constraints	Operation date	PATIENTID
15	RAWDATA major surgery (historical, further constraints)	OPCS-4 code	ICD-10 diagnosis code	Further notes/constraints	Operation date	PATIENTID
16	RAWDATA major surgery (new)	OPCS-4 code	ICD-10 diagnosis code	Further notes/constraints	Operation date	PATIENTID
17	Prior tumour diagnosis	Statusofregistration	ICD-10 diagnosis code	Stage_best	Diagnosisdatebest	PATIENTID
18	Tumour diagnosis (Final)	Statusofregistration	ICD-10 diagnosis code	Stage_best	Diagnosisdatebest	PATIENTID
19	Patient vital status date	Vitalstatus	ICD-10 Underlying cause of death		Vitalstatusdate	PATIENTID
20	RAWDATA holistic needs assessment record	HNA point of pathway : HNA offered : HNA staff role	Primary diagnosis	Laterality	Date of HNA	PATIENTID
21	RAWDATA staging	Inferred best stage	ICD-10 diagnosis code	T/N/M components	Collected stage date	PATIENTID
22	CWT First Seen	Source of referral	Categorisation of TWW, screening and consultant upgrade cases, where relevant	Suspected cancer referral type	Date first seen	NHSNUMBER
23	HES diagnostic event	OPCS-4 code	Description	BX/LD	Operation date	NHSNUMBER

EVENT_TYPE	EVENT_DESC	EVENT_PROPERTY_1	EVENT_PROPERTY_2	EVENT_PROPERTY_3	EVENT_DATE	Linkage
24	RAWDATA personal care and support plan	PCSP point of pathway : PCSP offered : PCSP staff role	Primary diagnosis	Laterality	PCSP date	PATIENTID
25	RAWDATA end of treatment summary	eots_date	Primary diagnosis	Laterality		PATIENTID
50	Skeleton Tumour creation	E_base_record type (COSD = England, CANISC = Wales)	ICD-10 diagnosis code		Diagnosisdate	PATIENTID
51	Diagnosis reported in COSD	Number of times reported	ICD-10 diagnosis code	E_base_record type	Diagnosisdate	NHSNUMBER
52	CWT estimated diagnosis date	CWT First Treatment Flag	CWT recorded primary diagnosis (ICD)	CWT Cancer Treatment Event Type	Adjusted treat period start	NHSNUMBER
53	HES inferred tumour	HES cancer group	ICD-10 diagnosis code		Episode start date	NHSNUMBER
54	COSD diagnosis submission	E_base_record primary diagnoses	ICD-10 diagnosis code (submission)		Diagnosis date (submission)	PATIENTID
55	RAWDATA biopsy record	Laterality	ICD-10 diagnosis code		Collected date/authorised date	PATIENTID
56	RAWDATA imaging record	Laterality	ICD-10 diagnosis code	Procedure_date - diagdate	Diagdate	PATIENTID
57	RAWDATA HNA diagnosis	Laterality	Primary diagnosis (ICD-10)		Diagdate	PATIENTID
101	Inferred diagnosis (54 only)	Event_property_1	ICD-10 diagnosis code	Cancer group	First recorded date	PATIENTID

*: Data dictionary: Primary cancer site for cancer faster diagnosis pathway (https://www.datadictionary.nhs.uk/attributes/primary_cancer_site_for_cancer_faster_diagnosis_pathway.html)

**: Data dictionary: Holistic needs assessment point of pathway for cancer (https://www.datadictionary.nhs.uk/attributes/holistic_needs_assessment_point_of_pathway_for_cancer.html?hl=holistic%2Cneeds%2Cassessment%2Cpoint%2Cpathway%2Ccancer)

Appendix 2 - List of Rapid Registration fields available

Table A2: AT_RAPID_TUMOUR: field list

COLUMN_NAME	DATA_TYPE	Notes
INDIVIDUALID	NUMBER(11,0)	Matches AT_RAPID_PATHWAY for each event with event_type=101
PATIENTID	NUMBER(19,0)	Matches AT_RAPID_PATHWAY for each event with event_type=101
NHSNUMBER	VARCHAR2(12 BYTE)	Matches AT_RAPID_PATHWAY for each event with event_type=101
TUMOUR_AVPID	NUMBER	Matches AT_RAPID_PATHWAY for each event with event_type=101
DIAGNOSISDATE	DATE	Matches AT_RAPID_PATHWAY for each event with event_type=101
TUMOUR_SITE	VARCHAR2(255 BYTE)	Matches AT_RAPID_PATHWAY for each event with event_type=101 (event_property_2)
BIRTHDATEBEST	DATE	Taken from Encore
SEX	VARCHAR2(255 BYTE)	Taken from Encore
POSTCODE	VARCHAR2(255 BYTE)	Taken from Encore
SURNAME	VARCHAR2(64 BYTE)	Taken from Encore
FORENAME	VARCHAR2(64 BYTE)	Taken from Encore
STAGE	VARCHAR2(255 BYTE)	Defined for selected cancer sites
ETHNICITY	VARCHAR2(255 BYTE)	Taken from Encore
FINAL_ROUTE	VARCHAR2(22 BYTE)	Final Route to Diagosis using an adapted version of the standard NCRAS methodology
QUINTILE_2019	VARCHAR2(26 BYTE)	Index of Multiple Deprivation quintile defined using the standard NCRAS methodology
CHRL_TOT_27_03	NUMBER	Charlson score defined using the standard NCRAS methodology

COLUMN_NAME	DATA_TYPE	Notes
TUMOUR_MORPHOLOGY	VARCHAR2(255 BYTE)	Tumour morphology as recorded in the COSD system
TUMOUR_PERFORMANCESTATUS	VARCHAR2(4 BYTE)	Patient performance status at time of diagnosis
BASISOFDIAGNOSIS	VARCHAR2(260 CHAR)	The basis of diagnosis (e.g. clinical; pathological; etc.)
LSOA11	VARCHAR2(27 BYTE)	LSOA of residence at time of diagnosis
SOURCE	VARCHAR2(7 BYTE)	The dataset used as the primary source for the RCRD registration
SOURCE_ID	VARCHAR2(64 BYTE)	The unique ID of the record used as the primary source for the RCRD registration

Appendix 3 - Cancer groups used for matching

Table A3: Rapid Registration ICD-10 tumour inclusion list

ICD	CANCER_GROUP	SCOPE	ICD	CANCER_GROUP	SCOPE
C00	Head & Neck	DQ & CD	C54	Gynae	DQ & CD
C01	Head & Neck	DQ & CD	C55	Gynae	DQ & CD
C02	Head & Neck	DQ & CD	C56	Gynae	DQ & CD
C03	Head & Neck	DQ & CD	C57	Gynae	DQ & CD
C04	Head & Neck	DQ & CD	C58	Gynae	DQ & CD
C05	Head & Neck	DQ & CD	C59	Other	DQ & CD
C06	Head & Neck	DQ & CD	C60	Urology	DQ & CD
C07	Head & Neck	DQ & CD	C61	Prostate	DQ & CD
C08	Head & Neck	DQ & CD	C62	Urology	DQ & CD
C09	Head & Neck	DQ & CD	C63	Urology	DQ & CD
C10	Head & Neck	DQ & CD	C64	Urology	DQ & CD
C11	Head & Neck	DQ & CD	C65	Urology	DQ & CD
C12	Head & Neck	DQ & CD	C66	Urology	DQ & CD
C13	Head & Neck	DQ & CD	C67	Urology	DQ & CD
C14	Head & Neck	DQ & CD	C68	Urology	DQ & CD
C15	O-G	DQ & CD	C69	Brain & CNS	DQ & CD
C16	O-G	DQ & CD	C70	Brain & CNS	DQ & CD
C17	Upper GI	DQ & CD	C71	Brain & CNS	DQ & CD
C18	Colorectal	DQ & CD	C72	Brain & CNS	DQ & CD
C19	Colorectal	DQ & CD	C73	Endocrine	DQ & CD
C20	Colorectal	DQ & CD	C74	Endocrine	DQ & CD
C21	Colorectal	DQ & CD	C75	Endocrine	DQ & CD
C22	Upper GI	DQ & CD	C76	Unknown Primary	DQ & CD
C23	Upper GI	DQ & CD	C77	Unknown Primary	DQ & CD
C24	Upper GI	DQ & CD	C78	Unknown Primary	DQ & CD
C25	Upper GI	DQ & CD	C79	Unknown Primary	DQ & CD
C26	Upper GI	DQ & CD	C80	Unknown Primary	DQ & CD
C27	Other	DQ & CD	C81	Haematological	DQ & CD
C28	Other	DQ & CD	C82	Haematological	DQ & CD
C29	Other	DQ & CD	C83	Haematological	DQ & CD
C30	Head & Neck	DQ & CD	C84	Haematological	DQ & CD
C31	Head & Neck	DQ & CD	C85	Haematological	DQ & CD
C32	Head & Neck	DQ & CD	C86	Haematological	DQ & CD
C33	Lung	DQ & CD	C87	Haematological	DQ & CD
C34	Lung	DQ & CD	C88	Haematological	DQ & CD
C35	Other	DQ & CD	C89	Haematological	DQ & CD
C36	Other	DQ & CD	C90	Haematological	DQ & CD
C37	Other	DQ & CD	C91	Haematological	DQ & CD

Scope: DQ = 'Included in this data quality document'; CD = 'Included in cancerdata.nhs.uk/covid-19/rcrd dashboard'

ICD	CANCER_GROUP	SCOPE	ICD	CANCER_GROUP	SCOPE
C38	Lung	DQ & CD	C92	Haematological	DQ & CD
C39	Lung	DQ & CD	C93	Haematological	DQ & CD
C40	Bone & ST	DQ & CD	C94	Haematological	DQ & CD
C41	Bone & ST	DQ & CD	C95	Haematological	DQ & CD
C42	Other	DQ & CD	C96	Haematological	DQ & CD
C43	Melanoma	DQ & CD	C97	Unknown Primary	DQ & CD
C44	NMSC	•	D05	Breast	DQ
C45	Lung	DQ & CD	D06	Gynae	•
C46	Bone & ST	DQ & CD	D09	Urology	DQ
C47	Brain & CNS	DQ & CD	D32	Brain & CNS	DQ
C48	Gynae	DQ & CD	D33	Brain & CNS	DQ
C49	Bone & ST	DQ & CD	D35	Brain & CNS	DQ
C50	Breast	DQ & CD	D41	Urology	DQ
C51	Gynae	DQ & CD	D42	Brain & CNS	DQ
C52	Gynae	DQ & CD	D43	Brain & CNS	DQ
C53	Gynae	DQ & CD	D44	Brain & CNS	DQ

Scope: DQ = 'Included in this data quality document'; CD = 'Included in cancerdata.nhs.uk/covid-19/rcrd dashboard'

Appendix 4 - Alternative defining events

Several options were considered as to the defining events for the Rapid Registrations. Both standalone datasets, subsets of standalone datasets, and combined datasets were explored and their FNE and FPE figures quantified. A subset of these alternatives are presented below as a demonstration of the process but the majority of this exploratory work is out of scope for this document.

Candidates for diagnosis events from the three main datasets that are rapidly available and have nominally full coverage of cancer patients are shown below (SACT and RTDS were also examined but data is not presented). Of the three, the CWT data has the best FPE but the FNE is substantially higher than the COSD dataset. HES produced the worst results in both measures. A filtering process was applied to the standalone COSD data to remove apparently new diagnoses that were actually recurrences of prior tumours. This improved the FPE at a cost of increasing the FNE. We continue to test whether this process can be further refined to improve the combined FPE and FNE figures, and monitor changes in the underlying datasets that might also give new opportunities to do so.

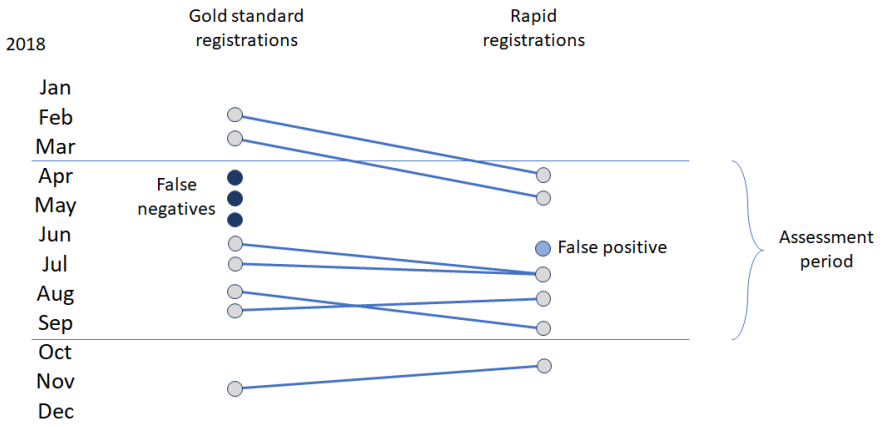
Table A4: Rapid Cancer Registrations: alternative defining events

Event	FPE	FNE
Event 52 - standalone CWT	7.6%	28.3%
Event 53 - standalone HES	13.2%	38.9%
Event 54 - standalone COSD	8.1%	15.8%
Event 101 (up to cas2106) - filtered COSD	5.2%	17.7%
Event 101 (cas2107) - filtered combined COSD/CWT	5.6%	16.4%
Event 101 (cas2108) - filtered combined COSD/CWT	5.1%	16.5%
Event 101 (cas2109) - filtered combined COSD/CWT	5.1%	16.6%
Event 101 (cas2110) - filtered combined COSD/CWT/HES	5.1%	14.7%
Event 101 (cas2111) - filtered combined COSD/CWT/HES	6.2%	13.4%
Event 101 (cas2112 to cas2202) - filtered combined COSD/CWT/HES and Death Certificates Only	5.3%	13.4%
Event 101 (cas2203 to cas2204) - filtered combined COSD/CWT/HES and Death Certificates Only	6.3%	12.2%
Event 101 (cas2205) - filtered combined COSD/CWT/HES and Death Certificates Only	6.1%	12.3%
Event 101 (cas2206) - filtered combined COSD/CWT/HES and Death Certificates Only	5.6%	12.5%
Event 101 (cas2207) - filtered combined COSD/CWT/HES and Death Certificates Only	6.0%	11.8%
Event 101 (cas2208 to cas2210) - filtered combined COSD/CWT/HES and Death Certificates Only	6.0%	11.6%
Event 101 (cas2211 to cas2302) - filtered combined COSD/CWT/HES and Death Certificates Only	6.1%	11.5%

Appendix 5 - Counts and error tabulations

Figure A1 shows an example for a very small dataset of how counts and error proportions are derived. This dataset has 10 Gold Standard Registrations and 7 Rapid Registrations overall (both indicated by the dots in the figure, with time running vertically over the course of 2018 and Gold Standard vs Rapid Registrations divided horizontally). Successful linkages between Gold Standard and Rapid Registrations are indicated by blue lines. False negatives and false positives are indicated. Only tumours in the 6-month assessment period are included in the tabulations below, although these can link to tumours outside the period as shown, and many-to-one linkages are also allowed. The false negative rate is therefore 3 in 7 and the false positive rate 1 in 6 below.

Figure A1: Illustration of counts and errors tabulation



Tables A5 and A6 below tabulate counts of Gold Standard and Rapid Registrations together with the numbers of false positive and false negative errors. When considering comparisons between figures the nature of the linkage and relationships displayed in the diagram above should be kept in mind.

Table A5: Counts and errors tabulation by cancer group

Cancer group	Gold Standard (GS) Registrations	Rapid Registrations	Difference	Percentage Rapid/GS	FPE	FNE
Brain & CNS	5589	5125	464	91.7%	691	1151
Breast	28920	27188	1732	94.0%	1497	1747
Colorectal	18957	17857	1100	94.2%	916	1672
Endocrine	1899	1680	219	88.5%	195	351
Gynae	9767	9446	321	96.7%	701	904
Haematological	13924	12525	1399	90.0%	802	2222
Head & Neck	5276	4932	344	93.5%	390	666
Lung	21652	20139	1513	93.0%	623	2025
Melanoma	8246	7695	551	93.3%	690	1045
O-G	6618	6482	136	97.9%	374	468
Prostate	27049	25243	1806	93.3%	312	2226
Bone & Soft Tissue	1140	1090	50	95.6%	367	407
Unknown Primary	3422	2661	761	77.8%	715	1478
Upper GI	9228	8765	463	95.0%	830	1344
Urology	16984	14762	2222	86.9%	917	2841

Table A6: Counts and errors tabulation by cancer site

Cancer site	Gold Standard (GS) Registrations	Rapid Registrations	Difference	Percentage Rapid/GS	FPE	FNE
C00	109	150	-41	137.6%	65	23
C01	645	470	175	72.9%	13	59
C02	604	618	-14	102.3%	17	85
C03	233	108	125	46.4%	4	64
C04	253	239	14	94.5%	9	30
C05	214	188	26	87.9%	8	31
C06	270	287	-17	106.3%	20	48
C07	235	284	-49	120.9%	100	50
C08	82	91	-9	111.0%	16	13
C09	913	775	138	84.9%	13	60
C10	151	233	-82	154.3%	11	29
C11	110	109	1	99.1%	6	11
C12	155	98	57	63.2%	1	10
C13	142	129	13	90.8%	11	21
C14	25	64	-39	256.0%	15	13
C15	3996	4322	-326	108.2%	126	217
C16	2622	2160	462	82.4%	248	251
C17	809	716	93	88.5%	152	230

Cancer site	Gold Standard (GS) Registrations	Rapid Registrations	Difference	Percentage Rapid/GS	FPE	FNE
C18	12425	11765	660	94.7%	668	1224
C19	994	955	39	96.1%	45	90
C20	4894	4493	401	91.8%	114	320
C21	644	644	0	100.0%	89	38
C22	2637	2539	98	96.3%	264	425
C23	472	475	-3	100.6%	30	55
C24	641	525	116	81.9%	28	84
C25	4518	4201	317	93.0%	138	477
C26	151	309	-158	204.6%	218	73
C30	162	155	7	95.7%	25	26
C31	92	64	28	69.6%	5	25
C32	881	870	11	98.8%	51	68
C33	13	12	1	92.3%	1	3
C34	20195	18768	1427	92.9%	549	1843
C37	167	86	81	51.5%	11	56
C38	72	356	-284	494.4%	47	21
C39	NA	13	NA	NA%	4	NA
C40	119	106	13	89.1%	11	25
C41	116	144	-28	124.1%	78	45
C43	8246	7695	551	93.3%	690	1045
C45	1205	904	301	75.0%	11	102
C46	68	43	25	63.2%	3	25
C47	26	14	12	53.8%	6	20
C48	285	453	-168	158.9%	141	72
C49	837	797	40	95.2%	275	312
C50	25096	24263	833	96.7%	1342	1424
C51	640	596	44	93.1%	56	77
C52	95	109	-14	114.7%	16	12
C53	1317	1325	-8	100.6%	53	76
C54	4095	3728	367	91.0%	111	177
C55	72	325	-253	451.4%	22	15
C56	2984	2571	413	86.2%	246	438
C57	269	313	-44	116.4%	37	36
C58	10	26	-16	260.0%	19	1
C60	303	292	11	96.4%	41	50
C61	27049	25243	1806	93.3%	312	2226
C62	1053	1073	-20	101.9%	86	70
C63	31	18	13	58.1%	7	24
C64	4844	4391	453	90.6%	277	732
C65	413	323	90	78.2%	23	82
C66	357	259	98	72.5%	13	116
C67	4472	5049	-577	112.9%	140	669
C68	95	57	38	60.0%	6	39
C69	370	328	42	88.6%	35	63
C70	20	45	-25	225.0%	4	1
C71	2258	2115	143	93.7%	135	190
C72	79	89	-10	112.7%	34	16
C73	1725	1514	211	87.8%	109	269
C74	116	116	0	100.0%	49	46
C75	58	50	8	86.2%	37	36
C76	94	224	-130	238.3%	121	53

Cancer site	Gold Standard (GS) Registrations	Rapid Registrations	Difference	Percentage Rapid/GS	FPE	FNE
C77	272	130	142	47.8%	63	125
C78	597	57	540	9.5%	25	324
C79	230	129	101	56.1%	53	120
C80	2229	2121	108	95.2%	453	856
C81	893	879	14	98.4%	17	59
C82	1206	1047	159	86.8%	15	140
C83	3146	2711	435	86.2%	36	328
C84	392	235	157	59.9%	15	121
C85	1372	1006	366	73.3%	64	308
C86	NA	102	NA	NA%	2	NA
C88	209	357	-148	170.8%	14	54
C90	2539	2203	336	86.8%	68	421
C91	2266	1901	365	83.9%	82	470
C92	1762	1647	115	93.5%	306	270
C93	23	190	-167	826.1%	27	1
C94	27	80	-53	296.3%	65	12
C95	50	68	-18	136.0%	12	13
C96	39	99	-60	253.8%	79	25
D05	3824	2925	899	76.5%	155	323
D09	4910	1277	3633	26.0%	262	914
D32	1397	1045	352	74.8%	92	425
D33	448	605	-157	135.0%	117	165
D35	464	547	-83	117.9%	188	112
D41	506	2023	-1517	399.8%	62	145
D42	143	16	127	11.2%	4	27
D43	267	265	2	99.3%	54	66
D44	117	56	61	47.9%	22	66

Appendix 6 - False negative errors and basis of diagnosis

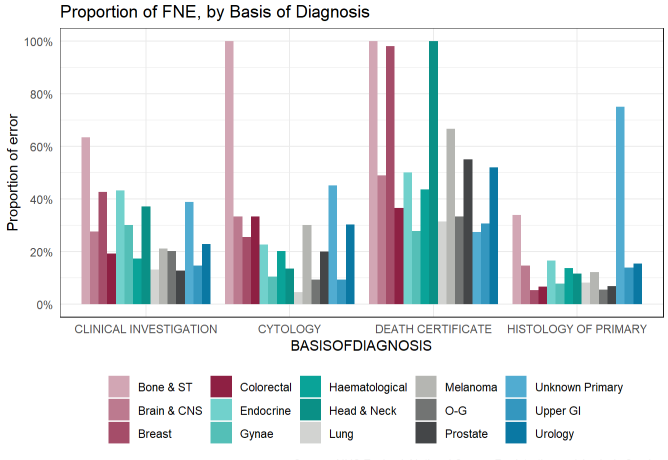
This appendix explores the reason for the overall age-dependence of the false negative error rate.

The most common methods of confirming a diagnosis (histology and cytology) account for the lowest proportion of false negatives (Figure A2). Where diagnosis comes from specific tumour markers, the Rapid Registrations are much more likely to “miss” the significant event or events. Patients diagnosed clinically (from imaging, consultation by a doctor but without a pathological sample being taken) are also more likely to be “missed” in the Rapid Registrations dataset.

Those patients for whom a diagnosis method cannot be determined (unknown) or died before they could be offered cancer treatment (death certificate), are most likely to be “missed” in the Rapid Registrations dataset. As Figure A3 indicates though, these account for a small proportion of those falsely omitted from the Rapid Registrations.

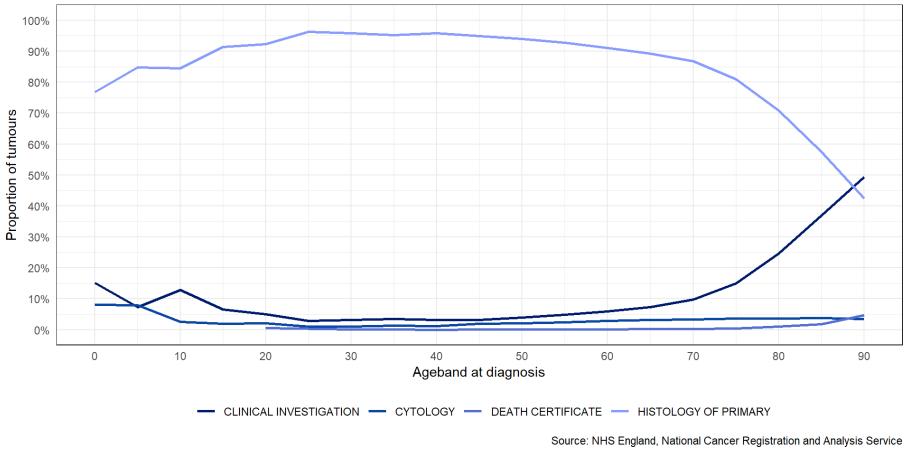
The marked reduction in the proportion of patients having their diagnosis confirmed from a pathological specimen (histology or cytology) explains the increase often observed at older ages in Figure A3, from the age of around 70, reflecting fewer patients having an invasive procedure performed on them as age increases. This is likely to be the reason behind the increasing false negative proportions by age observed overall and in most tumour groups (Figures 5 and 6).

Figure A2: The proportion of false negative Rapid Registrations by tumour group and basis of diagnosis, England, 2018



Source: NHS England, National Cancer Registration and Analysis Service

Figure A3: The proportion of false negative Rapid Registrations by method of diagnosis, England, 2018 (all tumour types combined)

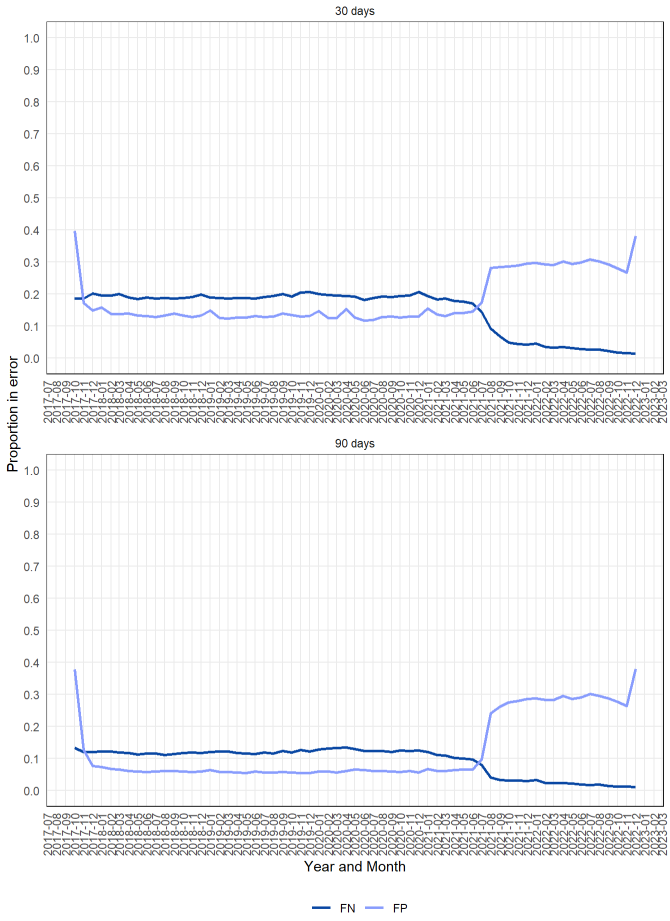


Source: NHS England, National Cancer Registration and Analysis Service

Appendix 7 - False positive and false negative proportion by month

Figure 18 shows the False Negative and False Positive error proportions by month for the broader matching criteria and a matching period of 90 and 30 days.

Figure A4: Monthly False Positive and False Negative proportions



Source: NHS England, National Cancer Registration and Analysis Service

Appendix 8 - Sensitivity testing of matching criteria

In this section, the sensitivity of the Rapid Registrations dataset is illustrated for different matching criteria.

As expected, the stricter the criteria about the timing of events, more errors (both false negative and false positive) are observed. Not including a match specification on tumour type (the second line of table 1) improves both matching criteria and demonstrates that approximately 40% of false positive tumours have a cancer diagnosis of some sort when the necessity of matching by tumour group is removed.

Table A7: Proportions of false positive and negative errors under alternative matching criteria

Tumour matching	Match within N days	False Negative %	False Positive %
Broader	90	11.5%	6.1%
Broader	60	13.2%	7.7%

Tumour matching	Match within N days	False Negative %	False Positive %
Broader	30	18.8%	13.4%
Broader	14	29.8%	25.1%
Broader	7	46.4%	42.9%
Broader	0	82.0%	80.7%
Narrow	90	19.4%	13.9%
None	90	10.0%	4.6%