Rapid Cancer Registration Dataset: data at 3rd July 2021 (CAS2107)

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 Public Health England (PHE) 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 16.4% of cases are missing and 5.6% 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.
- 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

Methodology

Proxy registration events (Rapid Registrations)

Data structures

Data Quality

How do the number of Rapid Registrations compare with Gold Standard Registrations?
Comparing the matching quality of Rapid Registrations
Sensitivity testing of matching criteria
Counts of events over time
Estimated completeness of Rapid Registrations and secondary datasets

Staging data in the Rapid Registrations dataset

TNM stage group 1-4
"Early" vs "Late" stage
Stage trends over time

Appendix 1 - List of pathway events

Appendix 2 - List of Rapid Registration fields available

Appendix 3 - Cancer groups used for matching

Appendix 4 - Alternative defining events

Appendix 5 - Counts and error tabulations

Appendix 6 - False negative errors and basis of diagnosis

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 Public Health 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

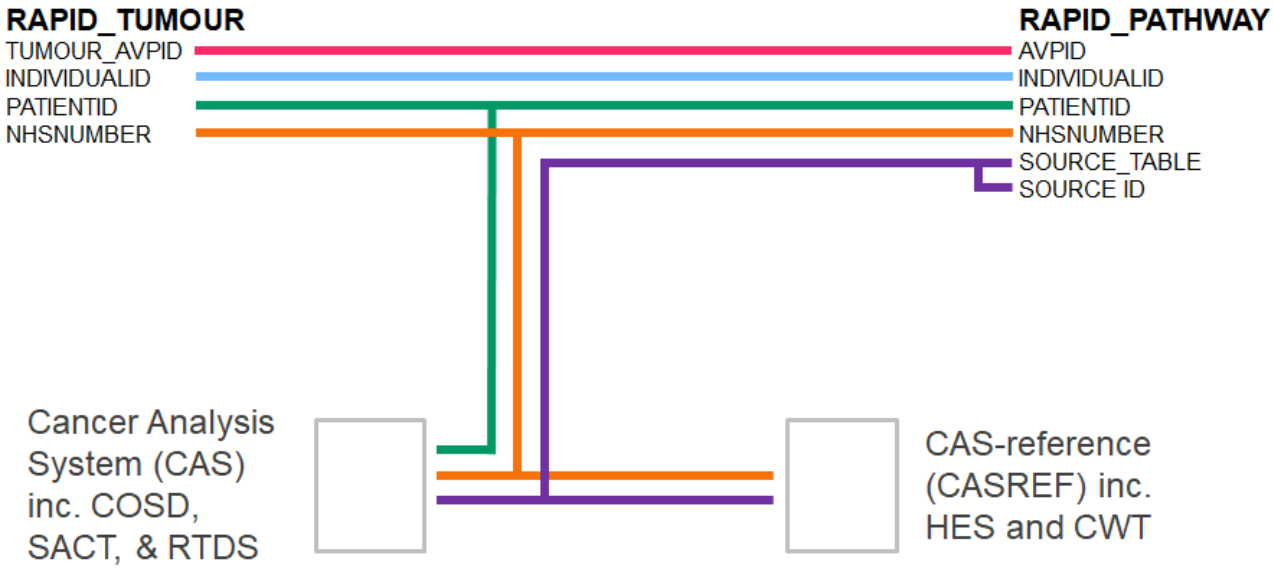
The rapid registration dataset consists of two tables:

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



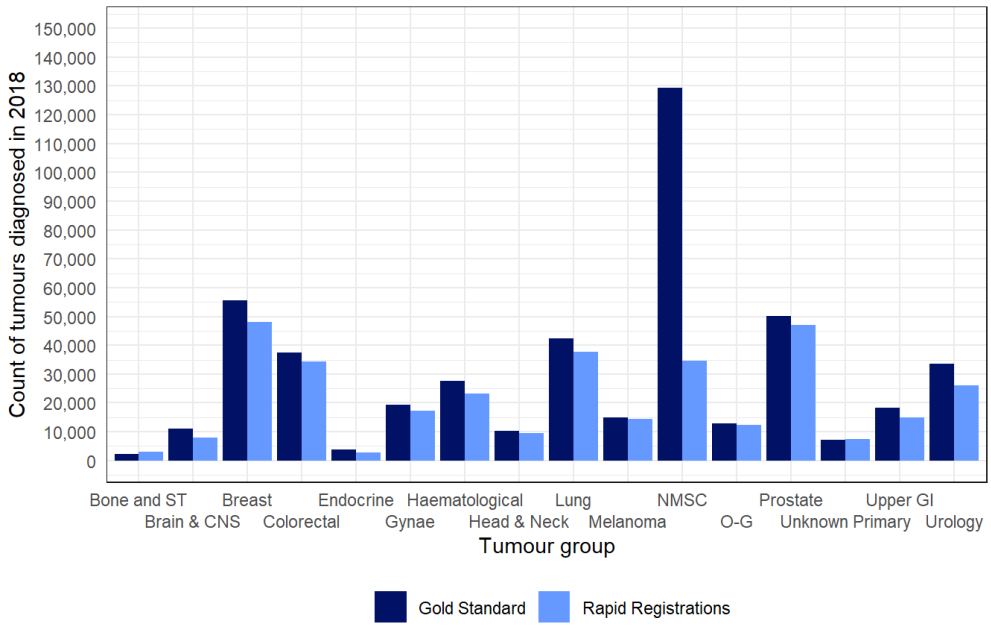
Data Quality

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.

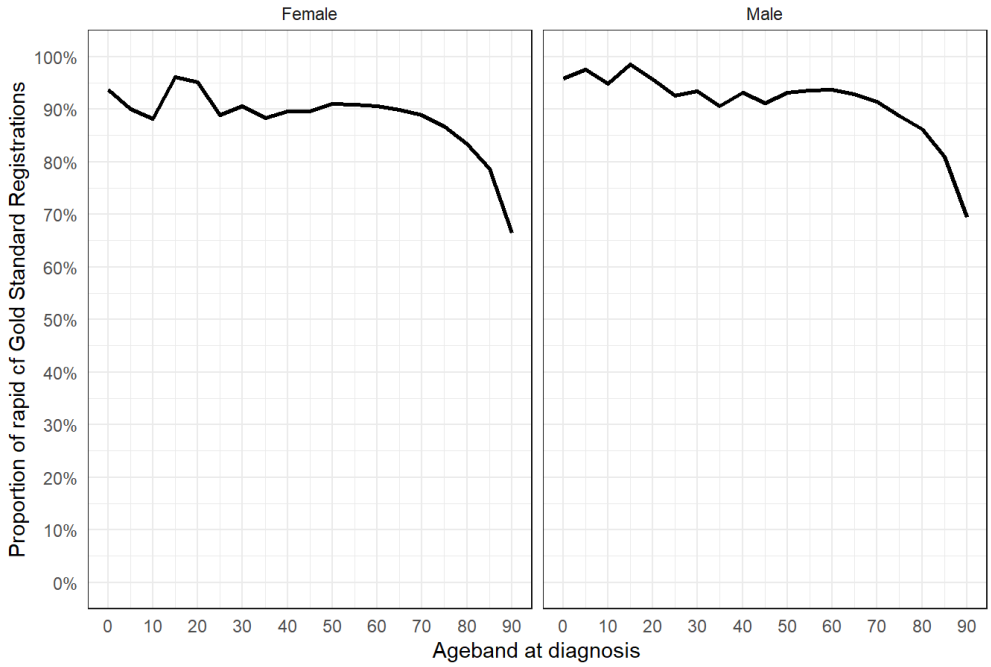
Figure 2: The number of cancer registrations by registration and tumour type, England, 2018



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

Comparing the matching quality of Rapid Registrations

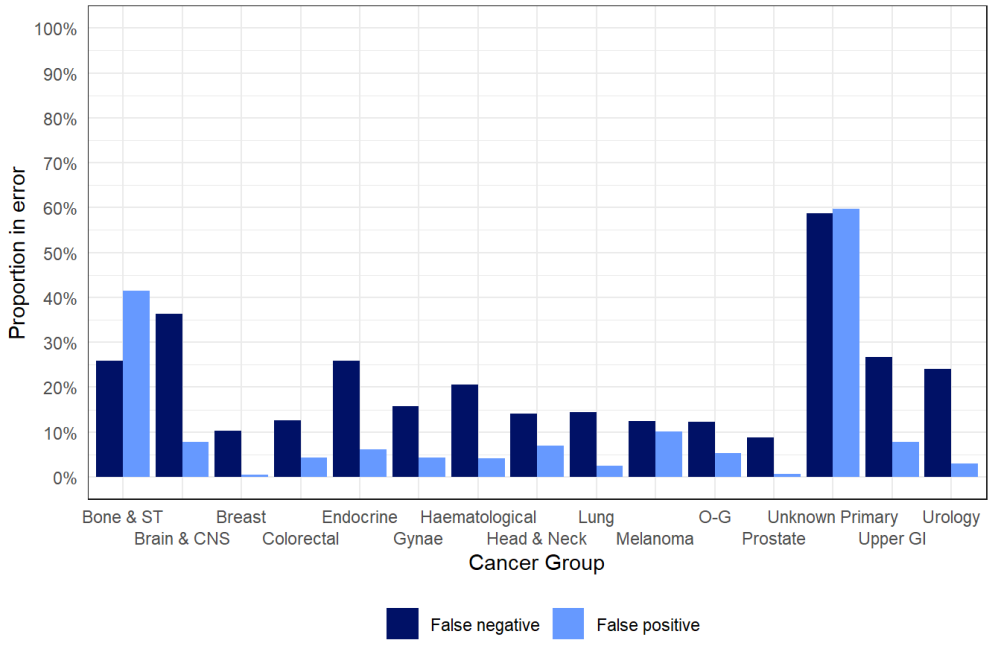
The quality of the Rapid Registrations was judged by comparing them against the gold-standard cancer registrations in the period April 2018 to September 2018. This period was chosen as available gold standard registration data was only finalised to December 2018 and a matching period of 90 days was allowed (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 cas2107 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 and prostate cancers are also least likely to be missing from the proxy dataset, whereas for brain and central nervous system (CNS), cancers of unknown primary, endocrine, bone and soft tissue, upper gastro-intestinal and urological tumours more than 25% of cancers are missed. Bone and soft tissue tumours, which have more false positives than false negatives, 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

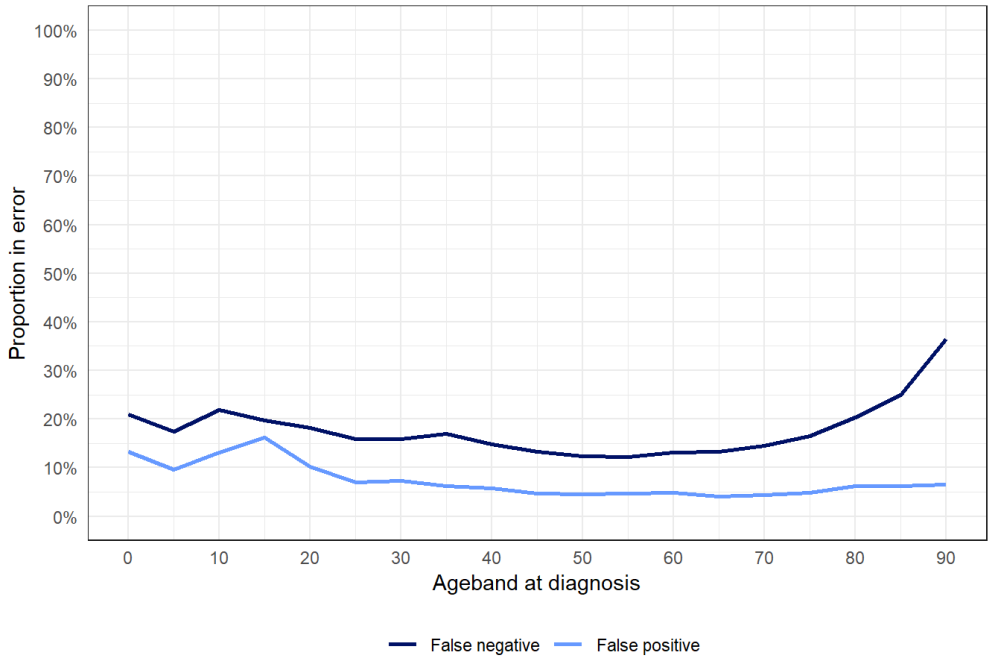


CNS: Central Nervous System; GI: Gastrointestinal; O-G: Oesophagogastric; ST: Soft Tissue
Source: Public Health England, National Cancer Registration and Analysis Service

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: Public Health 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.

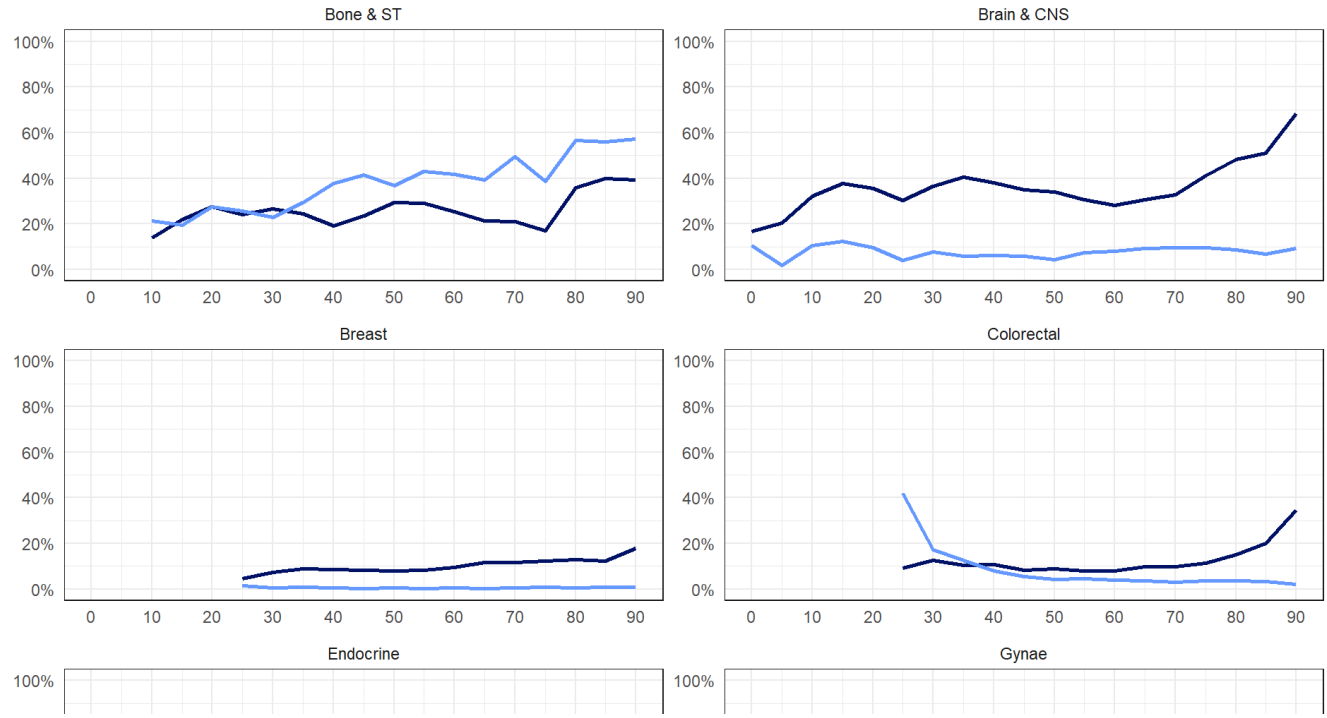
The patterns of false negative and false positive vary significantly by tumour group. Most groups have a higher proportion of false negatives than false positives at each age.

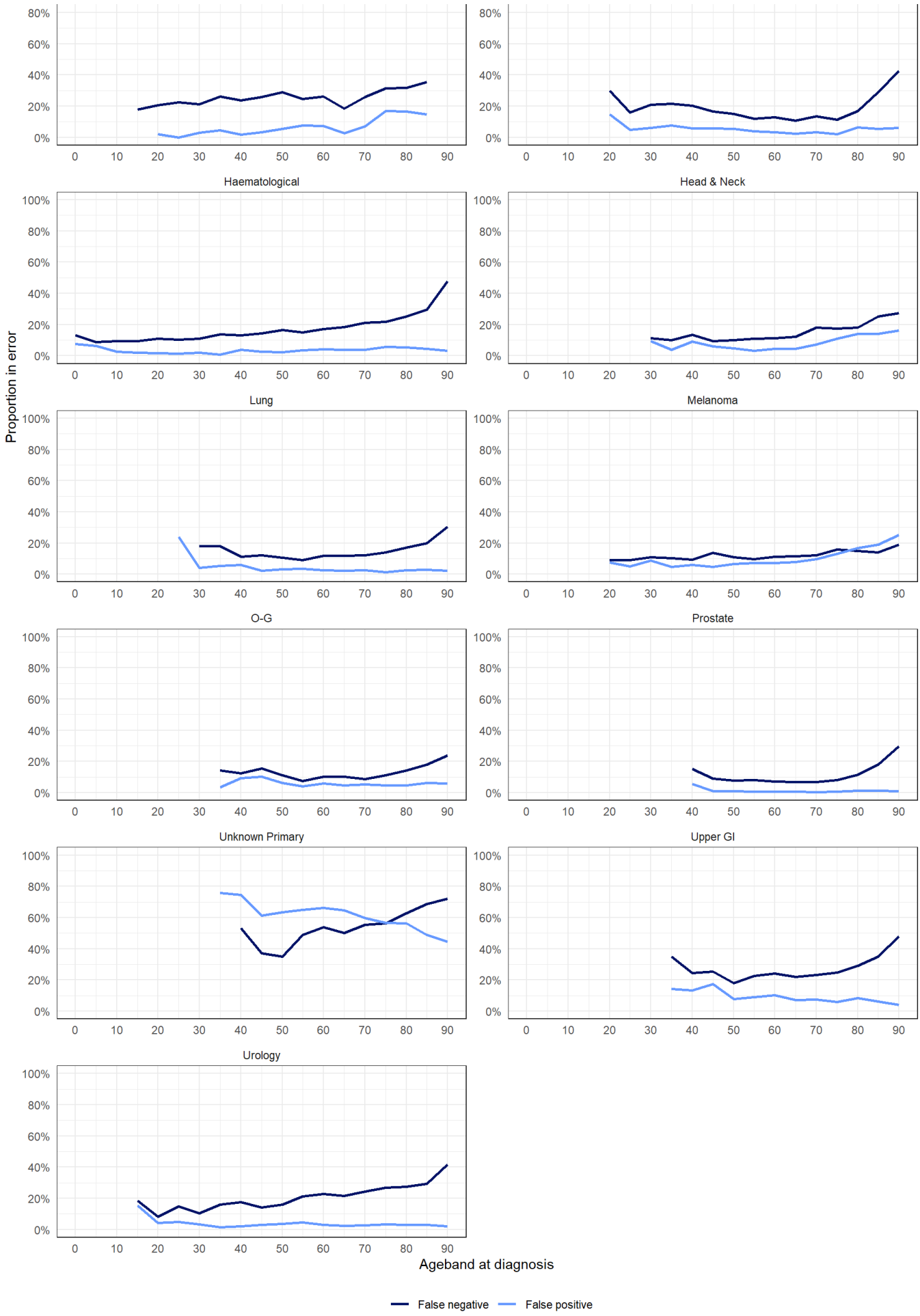
The proportion of false positives does not exhibit a trend by age for most tumour groups; the proportion rises with increasing age in the bone and soft tissue, head and neck groups and melanoma group and conversely falls with increasing age in the colorectal and unknown groups.

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

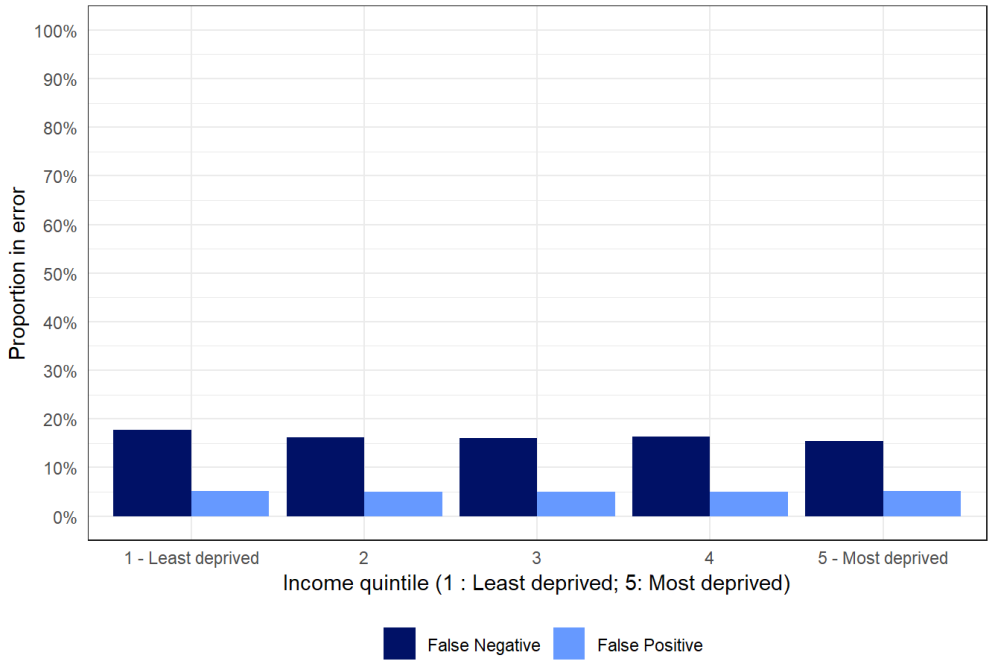




CNS: Central Nervous System; GI: Gastrointestinal; O-G: Oesophagogastric; ST: Soft Tissue
Source: Public Health England, National Cancer Registration and Analysis Service

The variation of the false positive and false negative errors with Income deprivation quintile is shown in figure 6. 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 6: False negative and false positive errors by income deprivation quintile

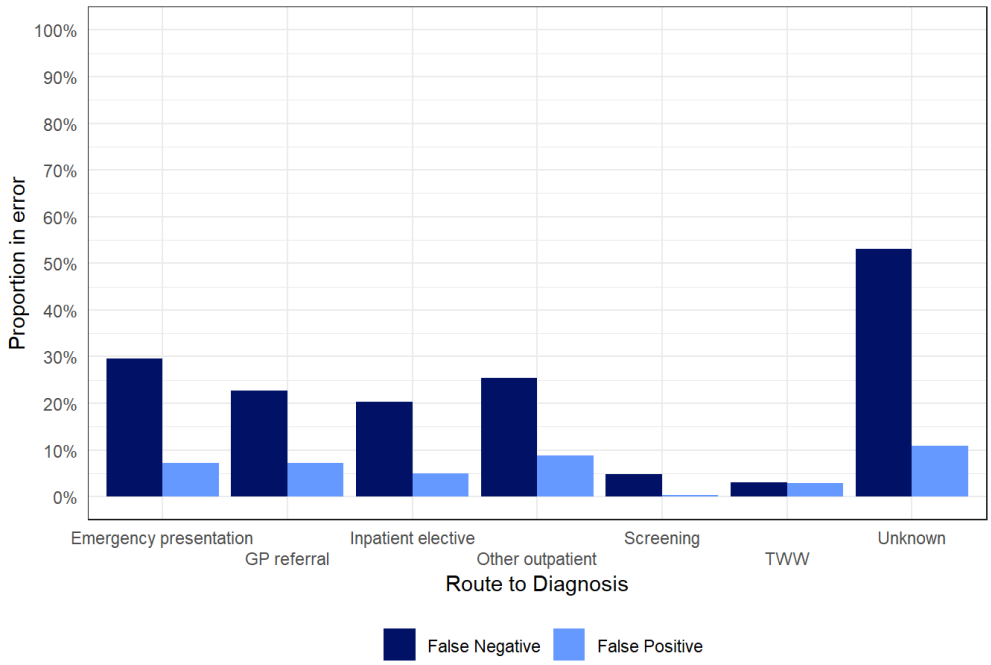


Source: Public Health England, National Cancer Registration and Analysis Service

Figure 7 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).

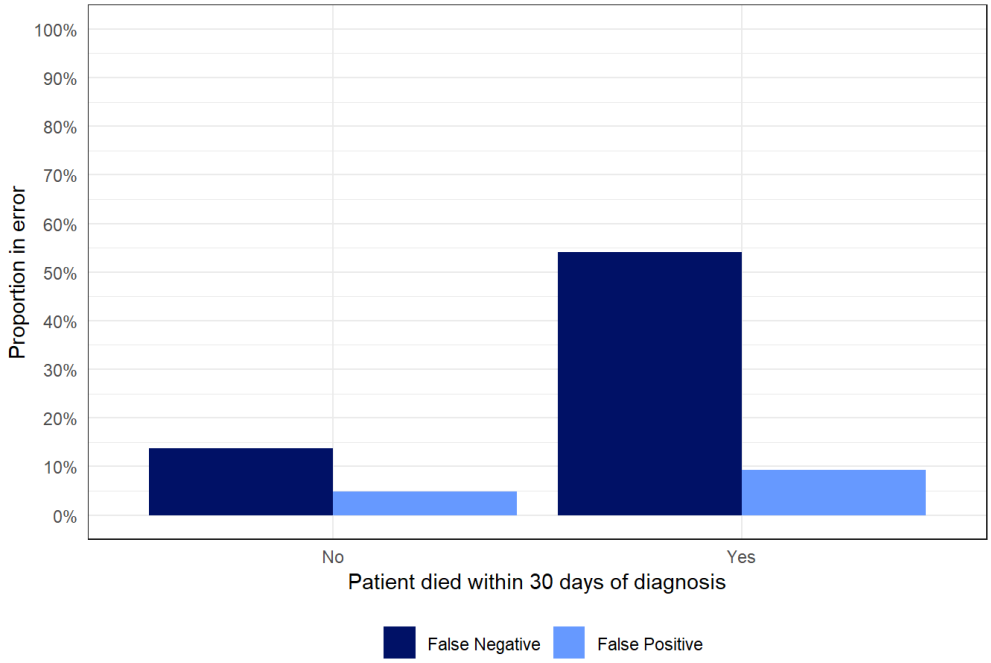
Figure 7: False negative and false positive errors by route to diagnosis



Source: Public Health England, National Cancer Registration and Analysis Service

Figure 8 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.

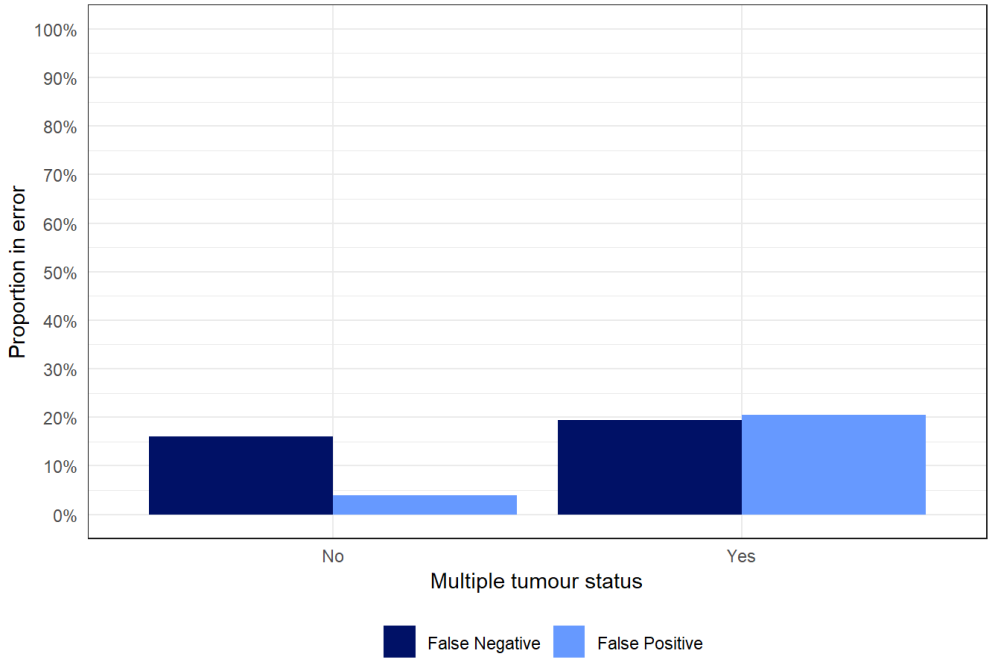
Figure 8: False negative and false positive errors by 30-day mortality



Source: Public Health England, National Cancer Registration and Analysis Service

Figure 9 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.

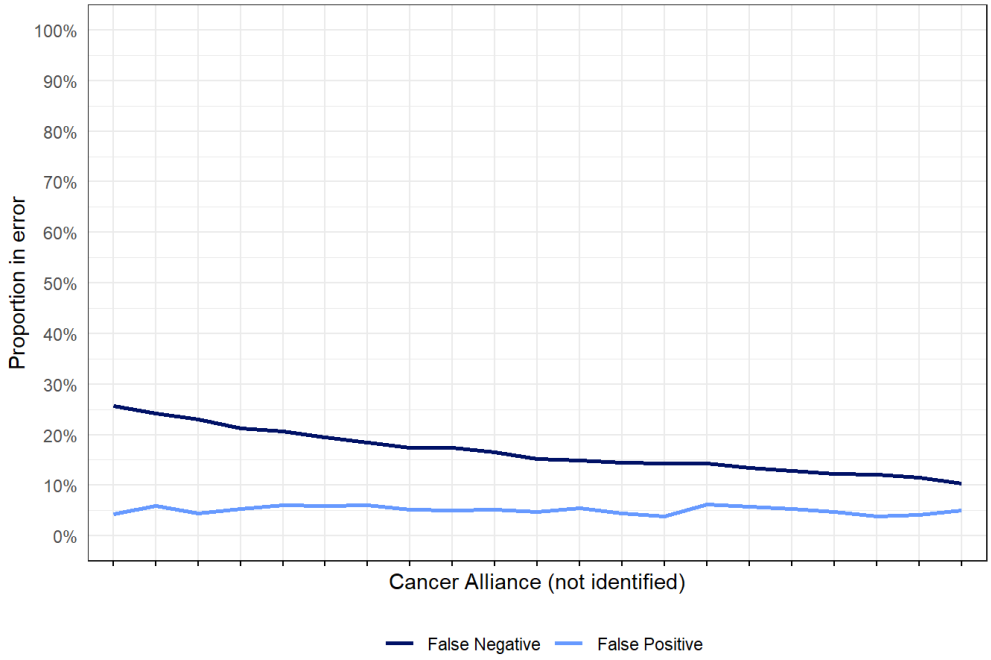
Figure 9: False negative and false positive errors by multiple tumour status



Source: Public Health England, National Cancer Registration and Analysis Service

Figure 10 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 10: False negative and false positive errors by Cancer Alliance

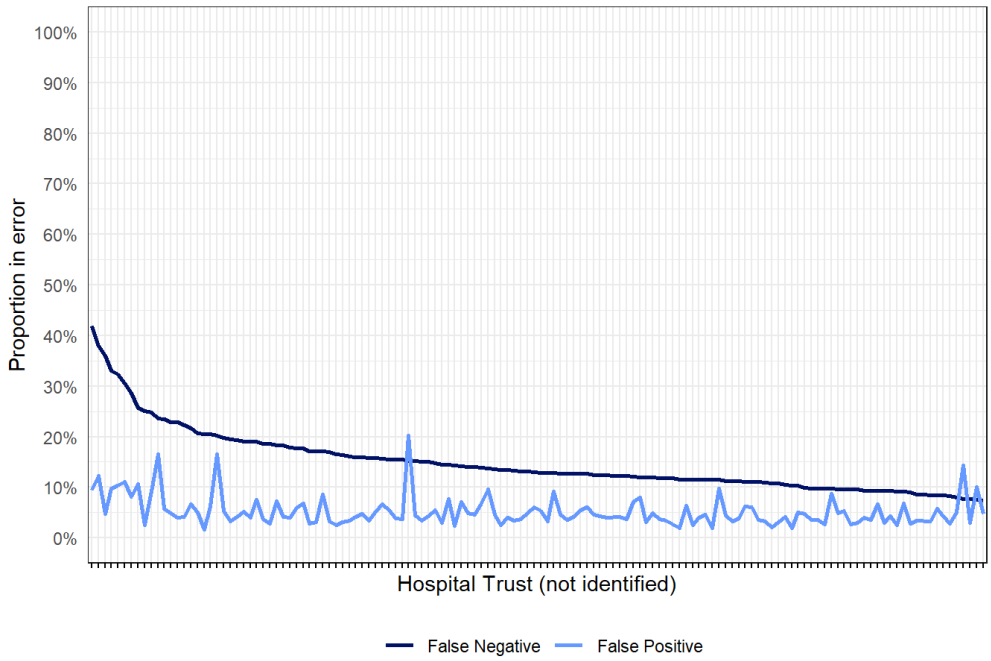


Source: Public Health England, National Cancer Registration and Analysis Service

Figures 11 and 12 below show the variation of false negative and false positive errors with the trust that diagnosed the tumour. Figure 11 shows the error proportion and figure 12 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.

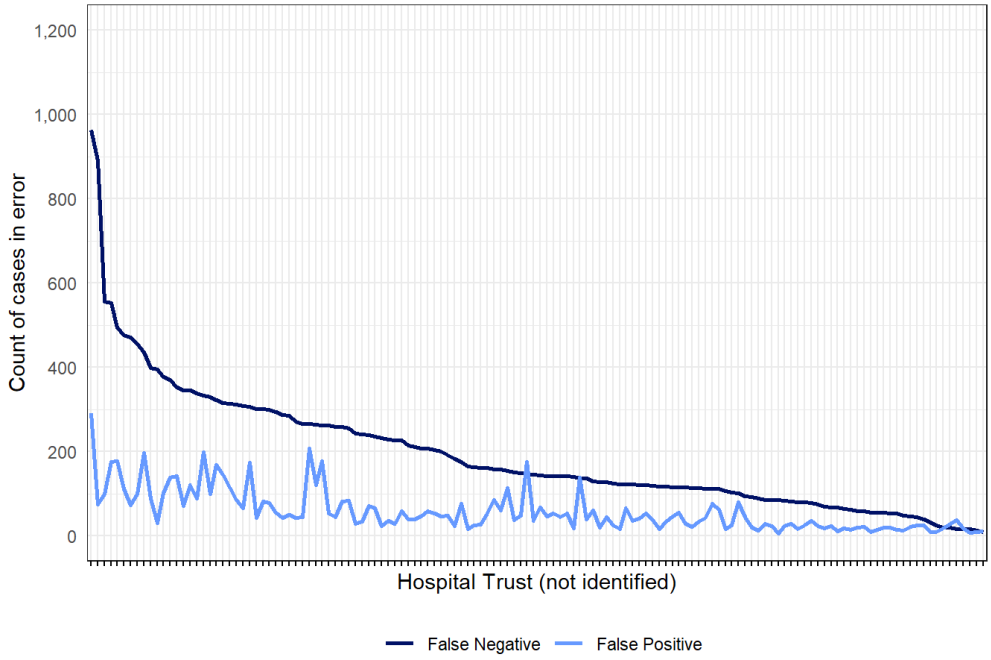
There is substantial variation in both false positive and false negative rates and counts. Some large trusts have several hundred or up to 1000 cases (over the six-month period under assessment).

Figure 11: False negative and false positive errors (proportion) by hospital trust



Source: Public Health England, National Cancer Registration and Analysis Service

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



Source: Public Health England, National Cancer Registration and Analysis Service

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

Tumour matching	Match within N days	False Negative %	False Positive %
Broader	90	16.4%	5.6%
Broader	60	17.7%	7.1%
Broader	30	22.7%	12.4%
Broader	14	33.1%	24.1%
Broader	7	49.1%	42.3%
Broader	0	83.0%	80.5%
Narrow	90	23.8%	13.4%
None	90	14.9%	3.4%

Counts of events over time

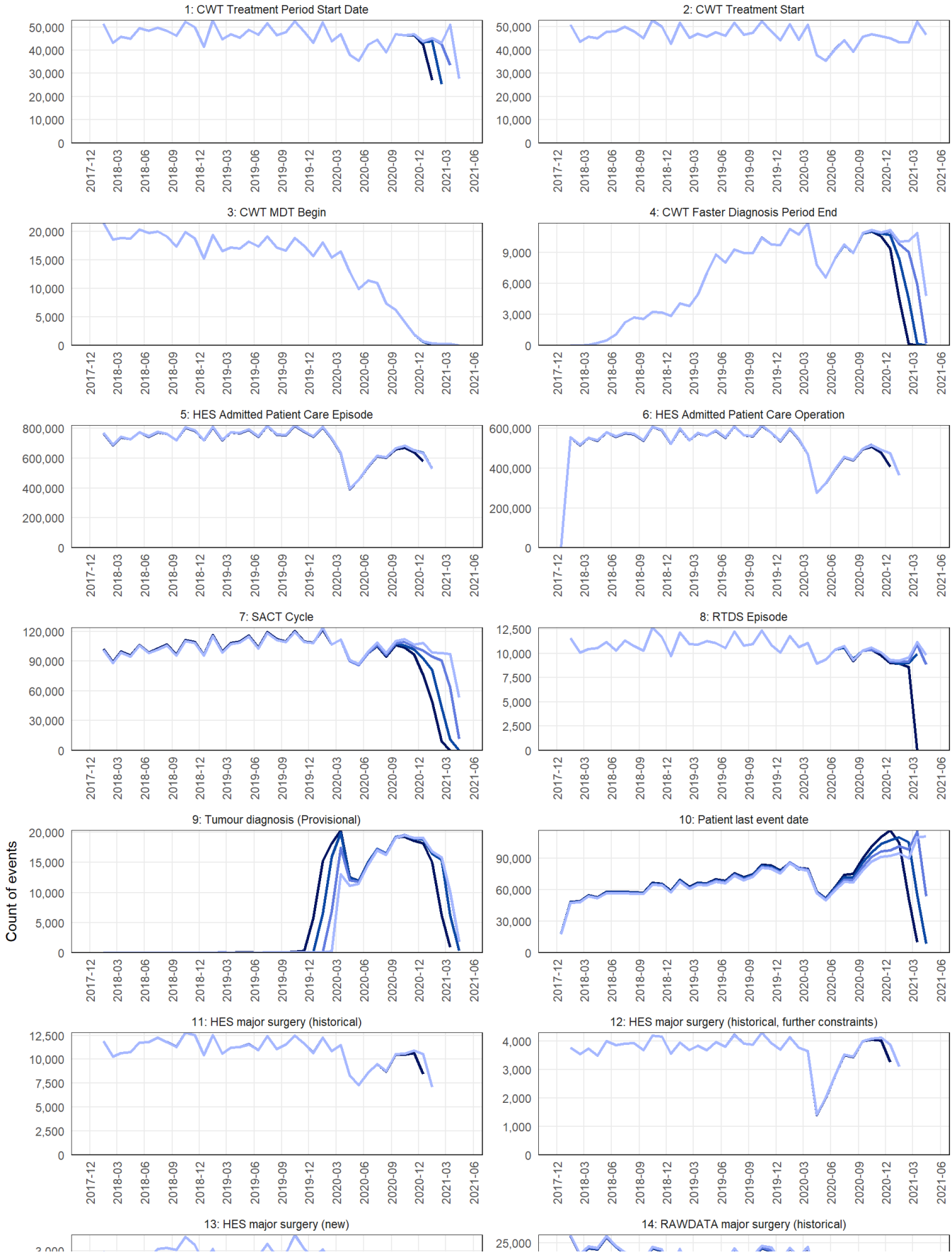
This section examines the population of events by chronological time and when they appear in successive analytical snapshots in the CAS. Figure 13 shows that most data items in the Rapid Registrations dataset are stable with respect to the snapshot month.

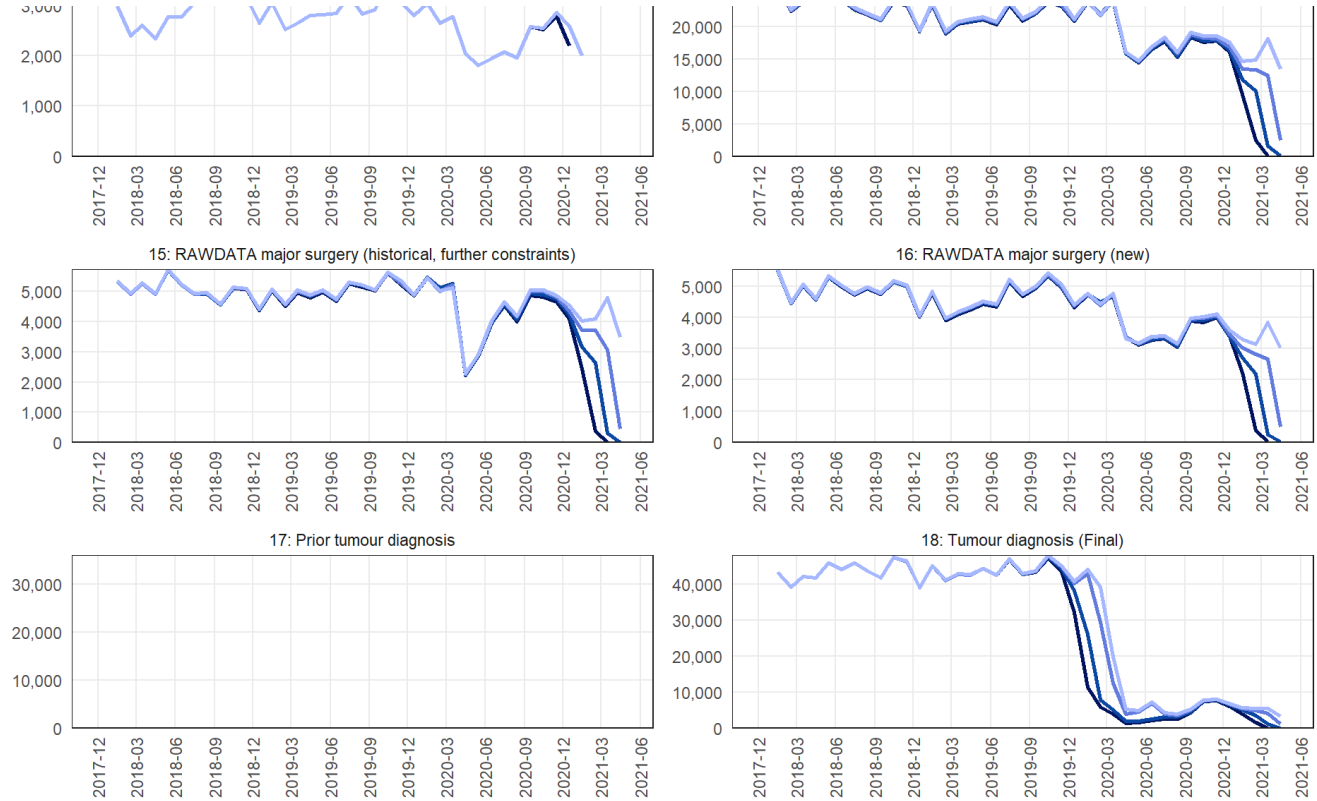
Specific comments about the events shown below are:

- 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.

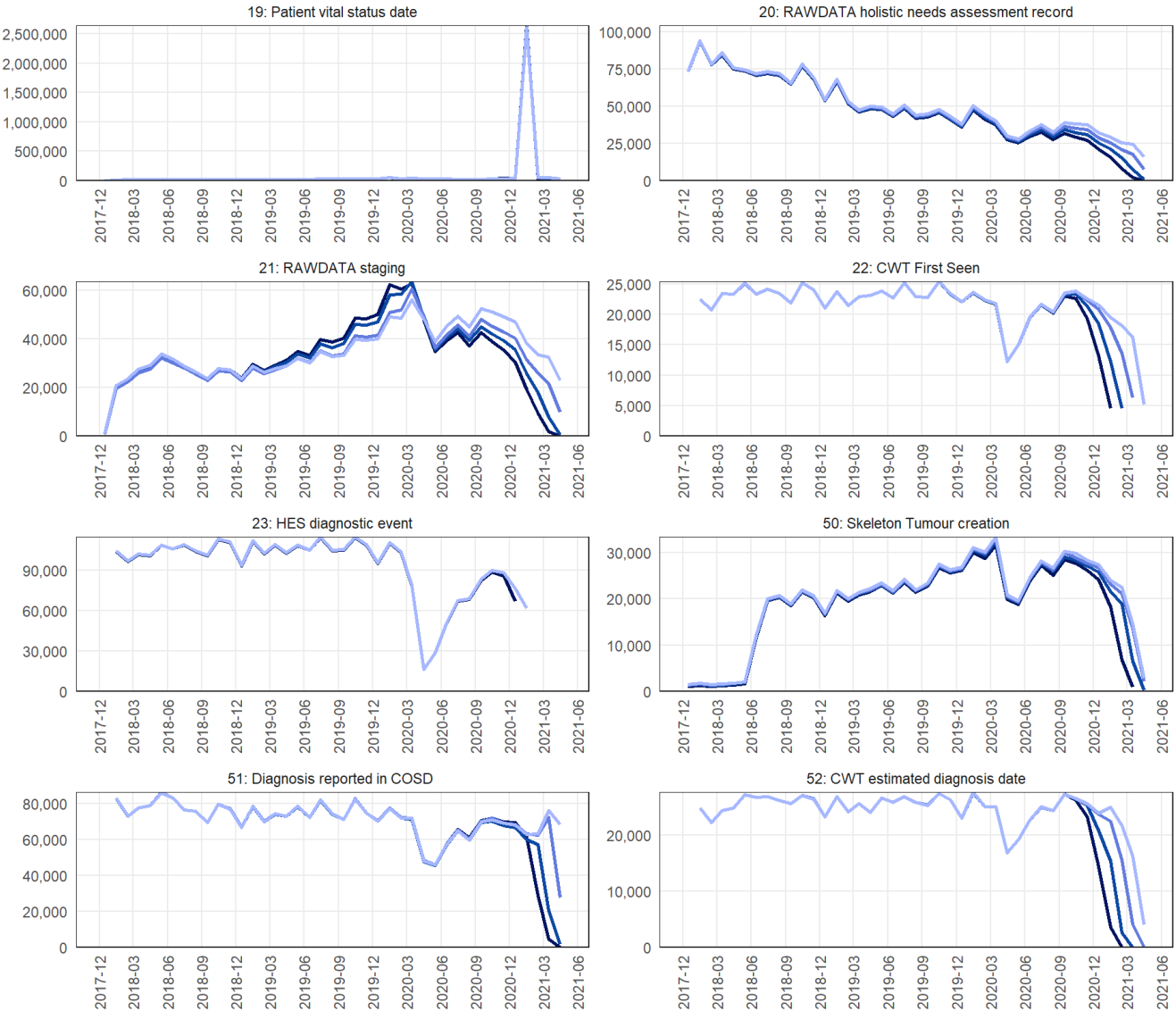
- The raw data used to populate events 21, 54, and 56 is subject to ongoing deduplication, this explains lower counts in earlier time periods for later snapshots.
- Between snapshots there is generally an increase in the Event 101-103 (Inferred diagnoses) counts, particularly for recent months as additional 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.

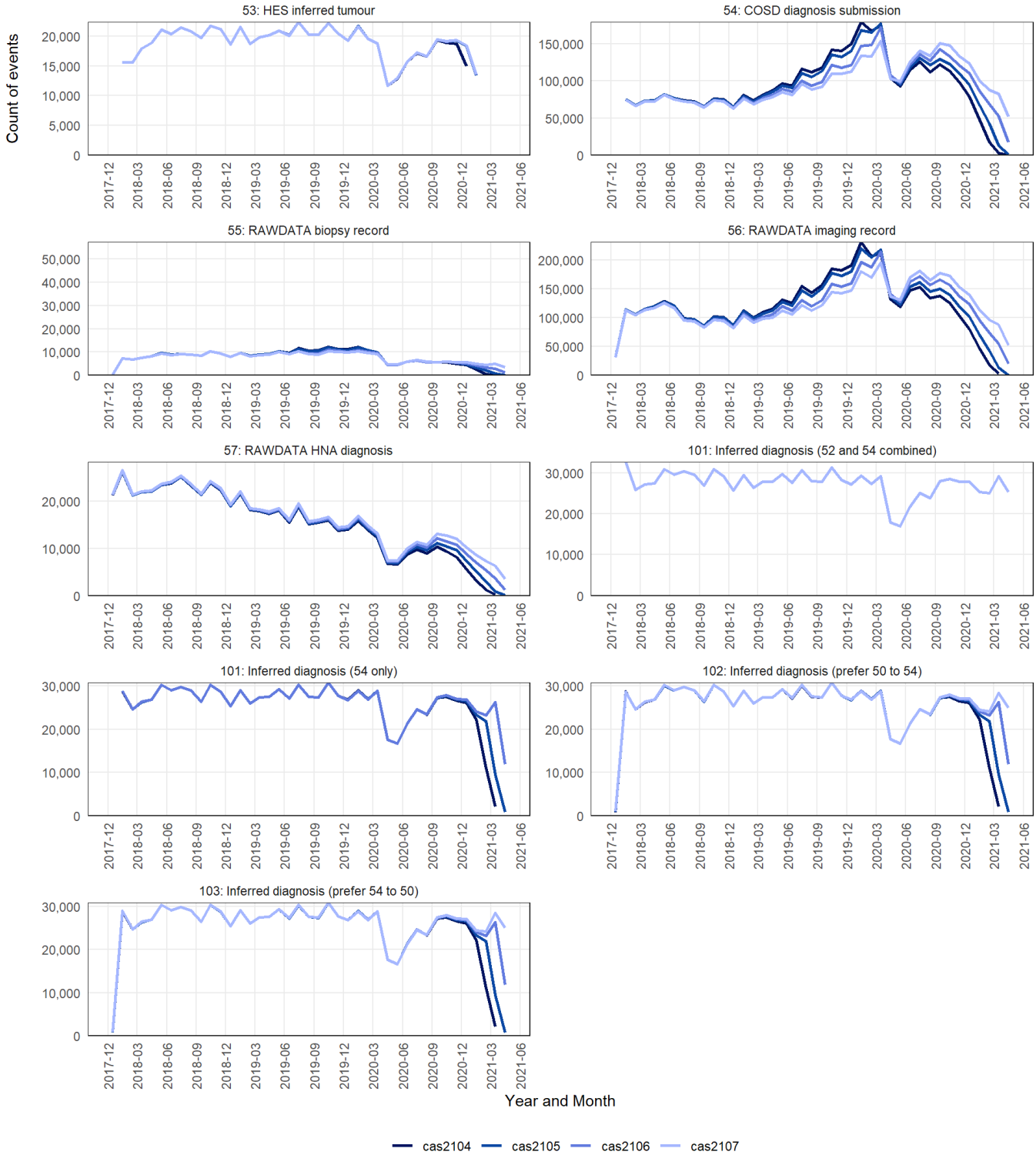
Figure 13: Population of data items to CAS snapshot





Source: Public Health England, National Cancer Registration and Analysis Service





Source: Public Health 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 cas2107

Data source	Latest usable	October 2020	November 2020	December 2020	January 2021	February 2021	March 2021	April 2021
Rapid Tumours (COSD)	April 2021	Complete	Complete	Complete	Complete	Complete	98%	91%
HES	December 2020	Complete	Complete	Complete	•	•	•	•
SACT	February 2021*	97%	95%	92%	88%	90%	•	•
RTDS	April 2021	Complete	98%	98%	98%	98%	93%	91%
CWT (TSD)	April 2021	Complete	Complete	Complete	Complete	Complete	Complete	Complete
CWT (TPSD)	March 2021	Complete	Complete	Complete	Complete	Complete	98%	62%

Note:
TSD = Treatment Start Date
TPSD = Treatment Period Start Date
* SACT data for January 2021 is below the 90% threshold and should be interpreted appropriately

Table 3: Number of outlier CCGs in COSD dataset in cas2107

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	0	135	135
2020-02	0	0	135	135
2020-03	0	1	134	135
2020-04	1	6	128	135
2020-05	1	4	130	135
2020-06	0	3	132	135
2020-07	0	0	135	135
2020-08	0	1	134	135
2020-09	1	0	134	135
2020-10	0	3	132	135
2020-11	0	0	135	135
2020-12	1	0	134	135
2021-01	1	0	134	135
2021-02	1	1	133	135
2021-03	1	2	132	135
2021-04	1	11	123	135

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 cas2107.

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	2308	25.7%	2787	31.1%
Bladder	2	1789	19.9%	1850	20.6%
Bladder	3	562	6.3%	843	9.4%
Bladder	4	254	2.8%	551	6.1%
Bladder	U	4062	45.3%	2944	32.8%
Breast	1	13308	31.9%	15708	37.7%
Breast	2	12773	30.7%	16041	38.5%
Breast	3	3113	7.5%	3566	8.6%
Breast	4	1094	2.6%	1741	4.2%
Breast	U	11384	27.3%	4616	11.1%
Colorectum	1	4933	15.9%	5395	17.4%
Colorectum	2	7033	22.6%	7547	24.3%
Colorectum	3	8266	26.6%	9218	29.7%
Colorectum	4	5116	16.5%	6908	22.2%
Colorectum	U	5706	18.4%	1986	6.4%
Kidney	1	2399	30.3%	3363	42.5%
Kidney	2	450	5.7%	543	6.9%
Kidney	3	1377	17.4%	1650	20.9%
Kidney	4	698	8.8%	1404	17.8%
Kidney	U	2983	37.7%	947	12.0%
Lung	1	6316	18.7%	6730	19.9%
Lung	2	2653	7.9%	2733	8.1%
Lung	3	7369	21.8%	7469	22.1%
Lung	4	14920	44.2%	16043	47.6%
Lung	U	2477	7.3%	760	2.3%
Lymphoma	1	911	8.0%	1716	15.0%
Lymphoma	2	949	8.3%	1569	13.7%
Lymphoma	3	1201	10.5%	1902	16.6%
Lymphoma	4	2654	23.2%	4634	40.5%
Lymphoma	U	5725	50.0%	1619	14.2%
Melanoma	1	6348	49.2%	8187	63.5%
Melanoma	2	2417	18.7%	2649	20.5%

Broad Cancer Group	Stage Group	Count (Rapid)	Percentage (Rapid)	Count (Gold Standard)	Percentage (Gold Standard)
Melanoma	3	440	3.4%	1041	8.1%
Melanoma	4	166	1.3%	285	2.2%
Melanoma	U	3523	27.3%	732	5.7%
Oesophagus	1	778	9.9%	442	5.6%
Oesophagus	2	1065	13.6%	955	12.2%
Oesophagus	3	2237	28.5%	2112	26.9%
Oesophagus	4	2002	25.5%	3084	39.3%
Oesophagus	U	1773	22.6%	1262	16.1%
Ovary	1	1119	24.7%	1314	29.0%
Ovary	2	236	5.2%	274	6.0%
Ovary	3	1170	25.8%	1570	34.6%
Ovary	4	696	15.3%	972	21.4%
Ovary	U	1317	29.0%	408	9.0%
Pancreas	1	347	5.2%	603	9.0%
Pancreas	2	633	9.4%	779	11.6%
Pancreas	3	749	11.1%	999	14.8%
Pancreas	4	2067	30.7%	3512	52.2%
Pancreas	U	2938	43.6%	841	12.5%
Prostate	1	11711	25.1%	16514	35.4%
Prostate	2	5503	11.8%	6740	14.4%
Prostate	3	10369	22.2%	11943	25.6%
Prostate	4	5625	12.1%	8030	17.2%
Prostate	U	13448	28.8%	3429	7.3%
Stomach	1	347	9.8%	334	9.4%
Stomach	2	282	7.9%	458	12.9%
Stomach	3	533	15.0%	692	19.5%
Stomach	4	1405	39.5%	1521	42.8%
Stomach	U	989	27.8%	551	15.5%
Uterus	1	4629	60.7%	5252	68.9%
Uterus	2	498	6.5%	523	6.9%
Uterus	3	724	9.5%	804	10.5%
Uterus	4	498	6.5%	518	6.8%
Uterus	U	1273	16.7%	525	6.9%

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	84.9%	3.6%	7.7%	5.1%	17.4%
2	4.0%	72.4%	15.7%	5.9%	8.9%
3	2.6%	10.9%	64.8%	3.9%	5.3%
4	1.3%	5.0%	5.9%	79.9%	4.8%
U	7.2%	8.1%	6.0%	5.1%	63.6%

b. breast (ICD-10 C50)

	Stage Group (Rapid)				
Stage Group (Gold Standard)	**1**	**2**	**3**	**4**	**Unknown**
1	89.5%	4.7%	1.4%	3.3%	27.4%
2	6.3%	88.9%	10.8%	14.1%	29.4%
3	0.5%	2.7%	80.9%	5.1%	5.1%
4	0.2%	0.8%	3.0%	73.1%	6.3%
U	3.5%	2.9%	3.9%	4.4%	31.7%

c. colorectum (ICD-10 C18-C20)

	Stage Group (Rapid)				
Stage Group (Gold Standard)	**1**	**2**	**3**	**4**	**Unknown**
1	85.1%	2.0%	1.7%	0.7%	15.5%
2	5.6%	85.7%	5.5%	1.2%	12.7%
3	6.5%	7.5%	85.4%	4.2%	19.2%
4	0.9%	2.8%	5.8%	92.9%	25.2%
U	1.8%	2.0%	1.7%	0.9%	27.4%

d. kidney (ICD-10 C64)

	Stage Group (Rapid)				
Stage Group (Gold Standard)	**1**	**2**	**3**	**4**	**Unknown**
1	91.4%	6.7%	3.1%	1.7%	36.4%
2	0.5%	78.2%	0.9%	0.7%	5.4%

	Stage Group (Rapid)				
Stage Group (Gold Standard)	**1**	**2**	**3**	**4**	**Unknown**
3	1.7%	6.4%	86.1%	3.9%	12.3%
4	0.4%	3.6%	5.7%	92.4%	22.0%
U	6.0%	5.1%	4.1%	1.3%	23.9%

e. lung (ICD-10 C33-C34)

	Stage Group (Rapid)				
Stage Group (Gold Standard)	**1**	**2**	**3**	**4**	**Unknown**
1	94.0%	6.1%	1.0%	0.4%	20.0%
2	2.6%	85.1%	1.8%	0.4%	5.2%
3	1.6%	4.9%	90.9%	1.2%	14.2%
4	1.2%	2.9%	5.4%	97.6%	37.6%
U	0.6%	0.9%	0.9%	0.4%	23.0%

f. melanoma (ICD-10 C43)

	Stage Group (Rapid)				
Stage Group (Gold Standard)	**1**	**2**	**3**	**4**	**Unknown**
1	94.6%	1.5%	4.1%	8.4%	60.0%
2	1.8%	79.4%	9.3%	15.7%	15.5%
3	1.9%	11.9%	80.9%	17.5%	7.0%
4	0.2%	1.6%	2.5%	54.8%	3.8%
U	1.5%	5.5%	3.2%	3.6%	13.7%

g. oesophagus (ICD-10 C15)

	Stage Group (Rapid)				
Stage Group (Gold Standard)	**1**	**2**	**3**	**4**	**Unknown**
1	40.2%	2.2%	0.3%	0.1%	5.5%
2	40.7%	42.1%	3.1%	0.7%	5.9%
3	10.2%	44.2%	57.8%	2.9%	11.8%
4	2.7%	5.9%	33.1%	86.4%	29.9%
U	6.2%	5.6%	5.6%	9.8%	46.9%

h. ovary (ICD-10 C56-C57)

	Stage Group (Rapid)				
Stage Group (Gold Standard)	**1**	**2**	**3**	**4**	**Unknown**
1	97.5%	6.4%	0.9%	0.3%	14.9%
2	0.4%	88.6%	0.5%	0.1%	4.1%

	Stage Group (Rapid)				
Stage Group (Gold Standard)	**1**	**2**	**3**	**4**	**Unknown**
3	0.8%	3.0%	91.9%	11.2%	30.4%
4	0.4%	0.4%	4.4%	84.2%	25.0%
U	1.0%	1.7%	2.3%	4.2%	25.6%

i. prostate (ICD-10 C61)

	Stage Group (Rapid)				
Stage Group (Gold Standard)	**1**	**2**	**3**	**4**	**Unknown**
1	86.3%	8.8%	3.9%	1.2%	40.6%
2	6.8%	83.9%	2.4%	0.9%	7.7%
3	4.3%	4.3%	87.1%	2.6%	15.0%
4	0.8%	0.7%	4.0%	93.4%	16.6%
U	1.9%	2.3%	2.6%	1.9%	20.2%

j. stomach (ICD-10 C16)

	Stage Group (Rapid)				
Stage Group (Gold Standard)	**1**	**2**	**3**	**4**	**Unknown**
1	65.4%	4.6%	0.6%	0.2%	8.9%
2	21.9%	64.9%	21.6%	1.7%	6.1%
3	5.5%	19.5%	57.0%	16.7%	8.0%
4	1.7%	7.1%	16.9%	78.9%	30.0%
U	5.5%	3.9%	3.9%	2.5%	47.0%

k. uterus (ICD-10 C54-C55)

	Stage Group (Rapid)				
Stage Group (Gold Standard)	**1**	**2**	**3**	**4**	**Unknown**
1	97.6%	11.4%	5.7%	7.4%	47.1%
2	0.6%	83.5%	1.2%	2.6%	4.6%
3	0.5%	1.8%	87.8%	7.2%	7.7%
4	0.2%	1.6%	2.3%	76.3%	8.2%
U	1.1%	1.6%	2.9%	6.4%	32.5%

l. pancreas (ICD-10 C25)

	Stage Group (Rapid)				
Stage Group (Gold Standard)	**1**	**2**	**3**	**4**	**Unknown**
1	73.5%	3.3%	1.1%	0.4%	10.6%
2	16.1%	76.1%	2.4%	0.4%	7.3%

	Stage Group (Rapid)				
Stage Group (Gold Standard)	**1**	**2**	**3**	**4**	**Unknown**
3	4.3%	11.5%	89.2%	0.6%	7.9%
4	3.5%	5.8%	6.0%	97.8%	47.5%
U	2.6%	3.2%	1.3%	0.8%	26.8%

m. lymphoma (ICD-10 C81-86, C88)

	Stage Group (Rapid)				
Stage Group (Gold Standard)	**1**	**2**	**3**	**4**	**Unknown**
1	90.7%	1.2%	0.6%	0.5%	15.0%
2	0.9%	93.8%	1.2%	0.4%	11.3%
3	0.3%	1.2%	90.1%	1.4%	13.4%
4	5.8%	2.3%	7.1%	93.2%	34.9%
U	2.3%	1.6%	1.1%	4.5%	25.3%

"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	4097	45.6%	4637	51.7%
Bladder	Late	816	9.1%	1394	15.5%
Bladder	Unknown	4062	45.3%	2944	32.8%
Breast	Early	26081	62.6%	31749	76.2%
Breast	Late	4207	10.1%	5307	12.7%
Breast	Unknown	11384	27.3%	4616	11.1%
Colorectum	Early	11966	38.5%	12942	41.7%
Colorectum	Late	13382	43.1%	16126	51.9%
Colorectum	Unknown	5706	18.4%	1986	6.4%
Kidney	Early	2849	36.0%	3906	49.4%
Kidney	Late	2075	26.2%	3054	38.6%
Kidney	Unknown	2983	37.7%	947	12.0%
Lung	Early	8969	26.6%	9463	28.1%
Lung	Late	22289	66.1%	23512	69.7%
Lung	Unknown	2477	7.3%	760	2.3%
Lymphoma	Early	1860	16.3%	3285	28.7%
Lymphoma	Late	3855	33.7%	6536	57.1%

Broad Cancer Group	Stage Group	Count (Rapid)	Percentage (Rapid)	Count (Gold Standard)	Percentage (Gold Standard)
Lymphoma	Unknown	5725	50.0%	1619	14.2%
Melanoma	Early	8765	68.0%	10836	84.0%
Melanoma	Late	606	4.7%	1326	10.3%
Melanoma	Unknown	3523	27.3%	732	5.7%
Oesophagus	Early	1843	23.5%	1397	17.8%
Oesophagus	Late	4239	54.0%	5196	66.1%
Oesophagus	Unknown	1773	22.6%	1262	16.1%
Ovary	Early	1355	29.9%	1588	35.0%
Ovary	Late	1866	41.1%	2542	56.0%
Ovary	Unknown	1317	29.0%	408	9.0%
Pancreas	Early	980	14.6%	1382	20.5%
Pancreas	Late	2816	41.8%	4511	67.0%
Pancreas	Unknown	2938	43.6%	841	12.5%
Prostate	Early	17214	36.9%	23254	49.8%
Prostate	Late	15994	34.3%	19973	42.8%
Prostate	Unknown	13448	28.8%	3429	7.3%
Stomach	Early	629	17.7%	792	22.3%
Stomach	Late	1938	54.5%	2213	62.2%
Stomach	Unknown	989	27.8%	551	15.5%
Uterus	Early	5127	67.3%	5775	75.8%
Uterus	Late	1222	16.0%	1322	17.3%
Uterus	Unknown	1273	16.7%	525	6.9%

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	83.3%	19.5%	26.2%
Late	9.1%	74.8%	10.1%
Unknown	7.6%	5.8%	63.6%

b. breast (ICD-10 C50)

	Stage Category (Rapid)		
Stage Category (Gold Standard)	**Early**	**Late**	**Unknown**
Early	94.8%	13.5%	56.8%
Late	2.0%	82.5%	11.5%

	Stage Category (Rapid)		
Stage Category (Gold Standard)	**Early**	**Late**	**Unknown**
Unknown	3.2%	4.0%	31.7%

c. colorectum (ICD-10 C18-C20)

	Stage Category (Rapid)		
Stage Category (Gold Standard)	**Early**	**Late**	**Unknown**
Early	88.9%	5.2%	28.2%
Late	9.1%	93.4%	44.4%
Unknown	2.0%	1.4%	27.4%

d. kidney (ICD-10 C64)

	Stage Category (Rapid)		
Stage Category (Gold Standard)	**Early**	**Late**	**Unknown**
Early	90.8%	3.5%	41.8%
Late	3.4%	93.3%	34.3%
Unknown	5.9%	3.2%	23.9%

e. lung (ICD-10 C33-C34)

	Stage Category (Rapid)		
Stage Category (Gold Standard)	**Early**	**Late**	**Unknown**
Early	95.0%	1.4%	25.2%
Late	4.3%	98.0%	51.8%
Unknown	0.7%	0.6%	23.0%

f. melanoma (ICD-10 C43)

	Stage Category (Rapid)		
Stage Category (Gold Standard)	**Early**	**Late**	**Unknown**
Early	92.1%	16.3%	75.5%
Late	5.2%	80.4%	10.8%
Unknown	2.6%	3.3%	13.7%

g. Oesophagus (ICD-10 C15)

	Stage Category (Rapid)		
Stage Category (Gold Standard)	**Early**	**Late**	**Unknown**
Early	59.7%	2.2%	11.4%
Late	34.4%	90.2%	41.7%
Unknown	5.9%	7.6%	46.9%

h. ovary (ICD-10 C56-C57)

	Stage Category (Rapid)		
Stage Category (Gold Standard)	**Early**	**Late**	**Unknown**
Early	97.3%	1.0%	19.0%
Late	1.5%	96.0%	55.4%
Unknown	1.1%	3.0%	25.6%

i. prostate (ICD-10 C61)

	Stage Category (Rapid)		
Stage Category (Gold Standard)	**Early**	**Late**	**Unknown**
Early	92.9%	4.8%	48.3%
Late	5.1%	92.9%	31.6%
Unknown	2.0%	2.3%	20.2%

j. stomach (ICD-10 C16)

	Stage Category (Rapid)		
Stage Category (Gold Standard)	**Early**	**Late**	**Unknown**
Early	79.3%	7.5%	15.0%
Late	15.9%	89.6%	38.0%
Unknown	4.8%	2.9%	47.0%

k. uterus (ICD-10 C54-C55)

	Stage Category (Rapid)		
Stage Category (Gold Standard)	**Early**	**Late**	**Unknown**
Early	97.9%	8.2%	51.6%
Late	1.0%	87.5%	15.9%
Unknown	1.1%	4.3%	32.5%

l. pancreas (ICD-10 C25)

	Stage Category (Rapid)		
Stage Category (Gold Standard)	**Early**	**Late**	**Unknown**
Early	83.1%	1.5%	17.9%
Late	14.0%	97.5%	55.4%
Unknown	3.0%	0.9%	26.8%

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

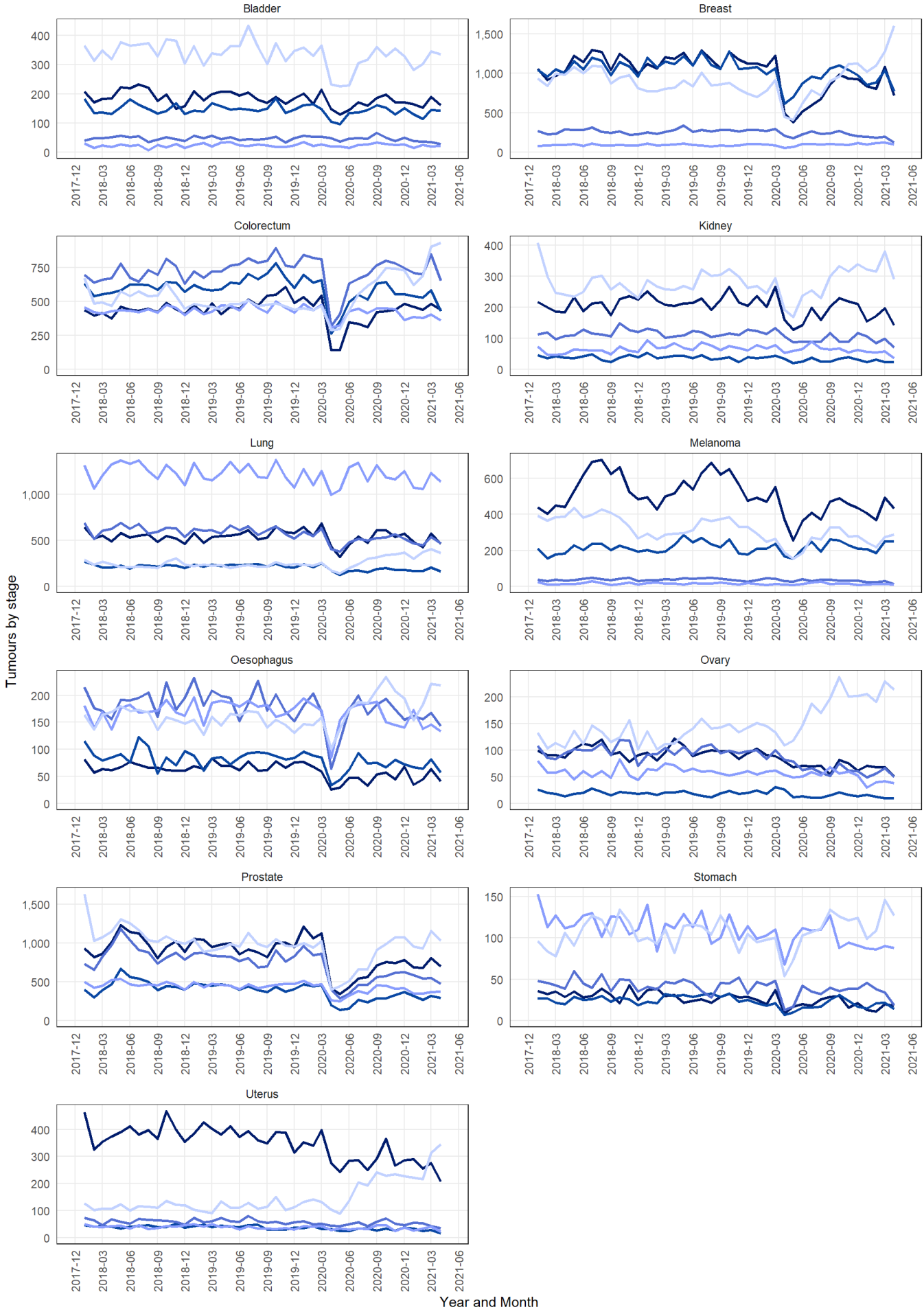
	Stage Category (Rapid)		
Stage Category (Gold Standard)	**Early**	**Late**	**Unknown**
Early	93.3%	1.1%	26.3%
Late	4.8%	95.4%	48.3%

	Stage Category (Rapid)		
Stage Category (Gold Standard)	**Early**	**Late**	**Unknown**
Unknown	1.9%	3.4%	25.3%

Stage trends over time

Figure 13 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 13: Stage trends over time

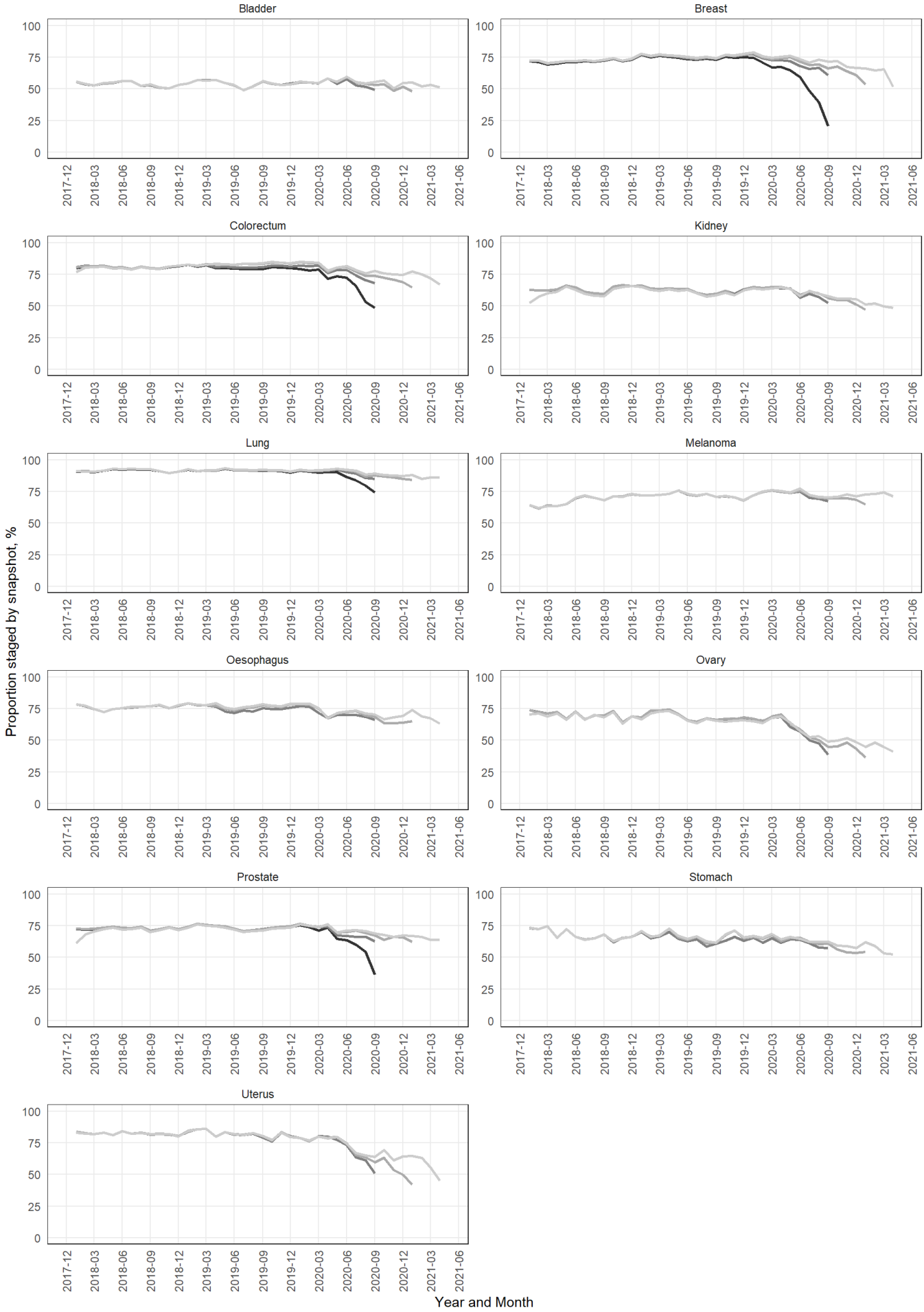


Source: Public Health England, National Cancer Registration and Analysis Service

Stage completeness by snapshot

Figure 14 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 14: Stage completeness by snapshot



Source: Public Health 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

EVENT_TYPE	EVENT_DESC	EVENT_PROPERTY_1	EVENT_PROPERTY_2	EVENT_PROPERTY_3	EVENT_DATE	Linkage
18	Tumour diagnosis (Final)	Statusofregistration	ICD-10 diagnosis code	Stage_best	Diagnosisdatebest	PATIENTID
19	Patient vital status date	Vitalstatus			Vitalstatusdate	PATIENTID
20	RAWDATA holistic needs assessment record	HNA point of pathway **	Primary diagnosis	Laterality	Date of HNA	PATIENTID
21	RAWDATA staging	Inferred best stage	ICD-10 diagnosis code	TNM components	Collected stage date	PATIENTID
22	CWT First Seen	REF_SOURCE	Categorisation of TWW, screening and consultant upgrade cases, where relevant	Suspected cancer referral type		NHSNUMBER
23	HES diagnostic event	OPCS-4 code	Description	BX/LD	Operation date	NHSNUMBER
50	Skeleton Tumour creation	E_base_record type	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 SITE_ICD10	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 diagonsis (ICD-10)		Diagdate	PATIENTID
101	Inferred diagnosis (54 only)	Event_property_1	ICD-10 diagnosis code	Cancer group	First recorded date	PATIENTID

*: https://www.datadictionary.nhs.uk/data_dictionary/attributes/p/prev/primary_cancer_site_for_cancer_faster_diagnosis_pathway_de.asp?shownav=0 (https://www.datadictionary.nhs.uk/data_dictionary/attributes/p/prev/primary_cancer_site_for_cancer_faster_diagnosis_pathway_de.asp?shownav=0)

**: https://www.datadictionary.nhs.uk/data_dictionary/attributes/h/ho/holistic_needs_assessment_point_of_pathway_for_cancer_de.asp?shownav=0 (https://www.datadictionary.nhs.uk/data_dictionary/attributes/h/ho/holistic_needs_assessment_point_of_pathway_for_cancer_de.asp?shownav=0)

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)	Income deprivation quintile defined using the standard NCRAS methodology
CHRL_TOT_27_03	NUMBER	Charlson score defined using the standard NCRAS methodology
TUMOUR_MORPHOLOGY	VARCHAR2(255 BYTE)	Tumour morphology as recorded in the COSD system

Appendix 3 - Cancer groups used for matching

Table A3: Rapid Registration ICD-10 tumour inclusion list

ICD	CANCER_GROUP	ICD	CANCER_GROUP
C00	Head & Neck	C54	Gynae
C01	Head & Neck	C55	Gynae
C02	Head & Neck	C56	Gynae
C03	Head & Neck	C57	Gynae
C04	Head & Neck	C58	Gynae
C05	Head & Neck	C59	Other
C06	Head & Neck	C60	Urology
C07	Head & Neck	C61	Prostate
C08	Head & Neck	C62	Urology
C09	Head & Neck	C63	Urology
C10	Head & Neck	C64	Urology

ICD	CANCER_GROUP	ICD	CANCER_GROUP
C11	Head & Neck	C65	Urology
C12	Head & Neck	C66	Urology
C13	Head & Neck	C67	Urology
C14	Head & Neck	C68	Urology
C15	O-G	C69	Brain & CNS
C16	O-G	C70	Brain & CNS
C17	Upper GI	C71	Brain & CNS
C18	Colorectal	C72	Brain & CNS
C19	Colorectal	C73	Endocrine
C20	Colorectal	C74	Endocrine
C21	Colorectal	C75	Endocrine
C22	Upper GI	C76	Unknown Primary
C23	Upper GI	C77	Unknown Primary
C24	Upper GI	C78	Unknown Primary
C25	Upper GI	C79	Unknown Primary
C26	Upper GI	C80	Unknown Primary
C27	Other	C81	Haematological
C28	Other	C82	Haematological
C29	Other	C83	Haematological
C30	Head & Neck	C84	Haematological
C31	Head & Neck	C85	Haematological
C32	Head & Neck	C86	Haematological
C33	Lung	C87	Haematological
C34	Lung	C88	Haematological
C35	Other	C89	Haematological
C36	Other	C90	Haematological
C37	Other	C91	Haematological
C38	Lung	C92	Haematological
C39	Lung	C93	Haematological
C40	Bone & ST	C94	Haematological
C41	Bone & ST	C95	Haematological
C42	Other	C96	Haematological
C43	Melanoma	C97	Unknown Primary
C44	NMSC	D05	Breast
C45	Lung	D06	Gynae
C46	Bone & ST	D09	Urology
C47	Brain & CNS	D32	Brain & CNS

ICD	CANCER_GROUP	ICD	CANCER_GROUP
C48	Gynae	D33	Brain & CNS
C49	Bone & ST	D35	Brain & CNS
C50	Breast	D41	Urology
C51	Gynae	D42	Brain & CNS
C52	Gynae	D43	Brain & CNS
C53	Gynae	D44	Brain & CNS

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.

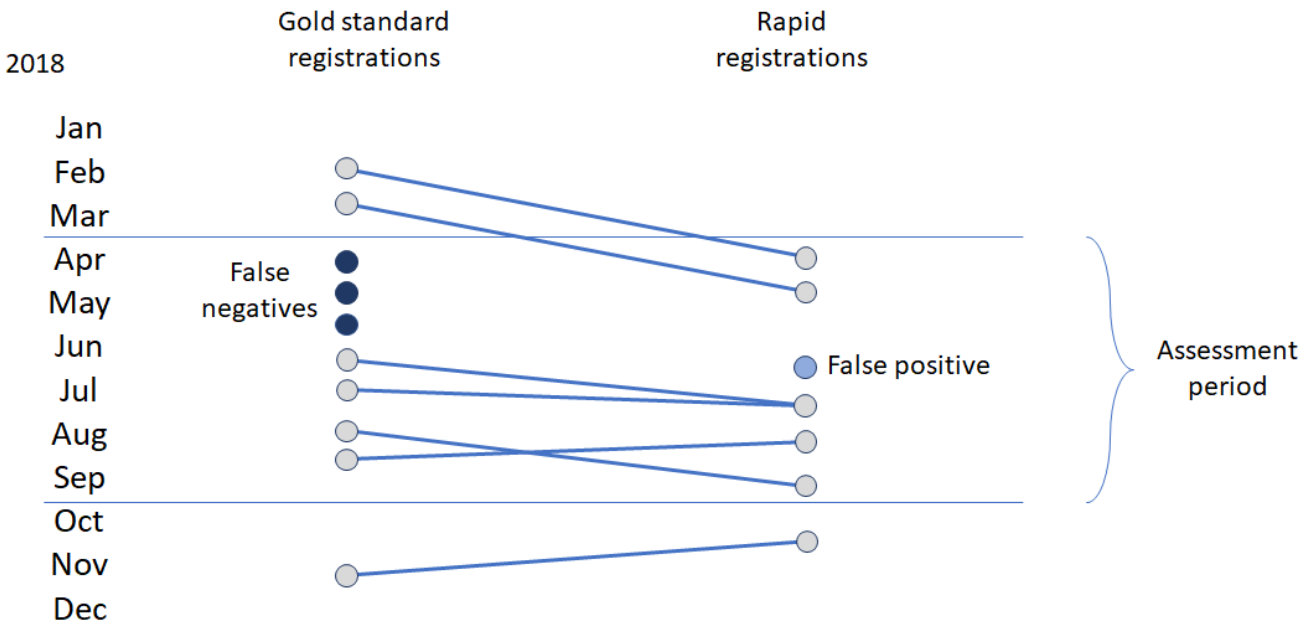
Table A4: Rapid Cancer Registrations: alternative defining events

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%

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	5429	3844	1585	70.8%	405	1897
Breast	28873	24666	4207	85.4%	233	1117
Colorectal	18892	16950	1942	89.7%	793	2011
Endocrine	1890	1466	424	77.6%	121	484
Gynae	9745	8503	1242	87.3%	421	1440
Haematological	13739	11441	2298	83.3%	523	2542
Head & Neck	5265	4811	454	91.4%	348	535
Lung	21522	18836	2686	87.5%	532	2763
Melanoma	8115	7717	398	95.1%	802	727
O-G	6615	6103	512	92.3%	338	792
Prostate	26873	24828	2045	92.4%	232	2344
Bone & Soft Tissue	1131	1427	-296	126.2%	604	285
Unknown Primary	3611	3640	-29	100.8%	2180	2118
Upper GI	9174	7392	1782	80.6%	640	2419
Urology	16880	13019	3861	77.1%	533	2876

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	144	-35	132.1%	59	16
C01	642	447	195	69.6%	9	50
C02	604	612	-8	101.3%	16	47
C03	233	104	129	44.6%	5	53

Cancer site	Gold Standard (GS) Registrations	Rapid Registrations	Difference	Percentage Rapid/GS	FPE	FNE
C04	251	238	13	94.8%	11	16
C05	214	186	28	86.9%	8	14
C06	270	280	-10	103.7%	18	39
C07	236	269	-33	114.0%	82	49
C08	81	88	-7	108.6%	15	13
C09	911	750	161	82.3%	14	61
C10	151	231	-80	153.0%	9	24
C11	110	104	6	94.5%	5	13
C12	155	99	56	63.9%	1	4
C13	142	123	19	86.6%	10	13
C14	24	61	-37	254.2%	13	12
C15	3996	4089	-93	102.3%	114	371
C16	2619	2014	605	76.9%	224	421
C17	804	650	154	80.8%	127	263
C18	12386	11135	1251	89.9%	599	1483
C19	992	816	176	82.3%	20	133
C20	4871	4379	492	89.9%	97	354
C21	643	620	23	96.4%	77	41
C22	2603	2120	483	81.4%	238	765
C23	473	422	51	89.2%	27	101
C24	642	478	164	74.5%	28	132
C25	4502	3582	920	79.6%	119	1039
C26	150	140	10	93.3%	101	119
C30	162	145	17	89.5%	19	27
C31	92	61	31	66.3%	4	24
C32	878	869	9	99.0%	50	60
C33	13	11	2	84.6%	1	3
C34	20069	17548	2521	87.4%	472	2525
C37	166	86	80	51.8%	11	58
C38	72	334	-262	463.9%	35	32
C39	NA	13	NA	NA%	4	NA
C40	118	111	7	94.1%	15	22
C41	115	197	-82	171.3%	122	35
C43	8115	7717	398	95.1%	802	727
C45	1202	844	358	70.2%	9	145
C46	68	46	22	67.6%	4	24

Cancer site	Gold Standard (GS) Registrations	Rapid Registrations	Difference	Percentage Rapid/GS	FPE	FNE
C47	26	15	11	57.7%	7	19
C48	283	377	-94	133.2%	107	87
C49	830	1073	-243	129.3%	463	204
C50	25059	22006	3053	87.8%	193	886
C51	641	497	144	77.5%	24	109
C52	92	94	-2	102.2%	10	11
C53	1318	1203	115	91.3%	36	168
C54	4091	3597	494	87.9%	79	268
C55	72	301	-229	418.1%	17	29
C56	2971	2122	849	71.4%	108	714
C57	267	290	-23	108.6%	23	51
C58	10	22	-12	220.0%	17	3
C60	302	294	8	97.4%	40	34
C61	26873	24828	2045	92.4%	232	2344
C62	1053	1025	28	97.3%	70	66
C63	29	17	12	58.6%	6	23
C64	4788	4006	782	83.7%	213	847
C65	409	303	106	74.1%	18	56
C66	356	235	121	66.0%	9	70
C67	4452	4811	-359	108.1%	117	485
C68	95	48	47	50.5%	5	21
C69	370	331	39	89.5%	36	58
C70	20	36	-16	180.0%	6	8
C71	2244	1839	405	82.0%	166	512
C72	78	75	3	96.2%	31	23
C73	1721	1358	363	78.9%	72	385
C74	115	67	48	58.3%	21	66
C75	54	41	13	75.9%	28	33
C76	94	553	-459	588.3%	456	74
C77	298	346	-48	116.1%	247	87
C78	682	272	410	39.9%	222	453
C79	287	373	-86	130.0%	278	193
C80	2250	2096	154	93.2%	977	1311
C81	894	844	50	94.4%	8	64
C82	1202	1031	171	85.8%	8	102
C83	3142	2610	532	83.1%	31	356

Cancer site	Gold Standard (GS) Registrations	Rapid Registrations	Difference	Percentage Rapid/GS	FPE	FNE
C84	386	220	166	57.0%	13	116
C85	1342	824	518	61.4%	41	391
C86	NA	94	NA	NA%	3	NA
C88	201	366	-165	182.1%	12	38
C90	2511	2002	509	79.7%	35	534
C91	2181	1759	422	80.7%	61	460
C92	1741	1244	497	71.5%	83	423
C93	23	149	-126	647.8%	8	4
C94	29	133	-104	458.6%	114	11
C95	50	38	12	76.0%	2	26
C96	37	127	-90	343.2%	104	17
D05	3814	2660	1154	69.7%	40	231
D09	4891	410	4481	8.4%	33	1060
D32	1322	705	617	53.3%	30	595
D33	413	478	-65	115.7%	62	193
D35	448	251	197	56.0%	30	230
D41	505	1870	-1365	370.3%	22	214
D42	138	6	132	4.3%	1	53
D43	260	85	175	32.7%	23	132
D44	110	23	87	20.9%	13	74

Appendix 6 - False negative errors and basis of diagnosis

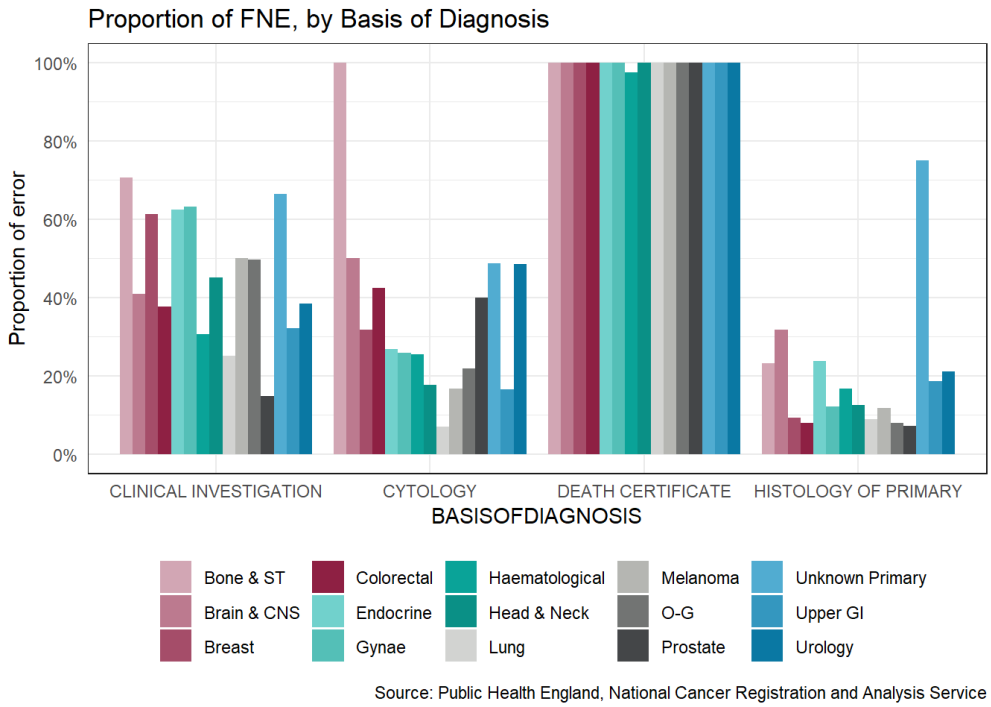
This appendix explores the reason for the overall age-dependence of the false negative error rate.

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.

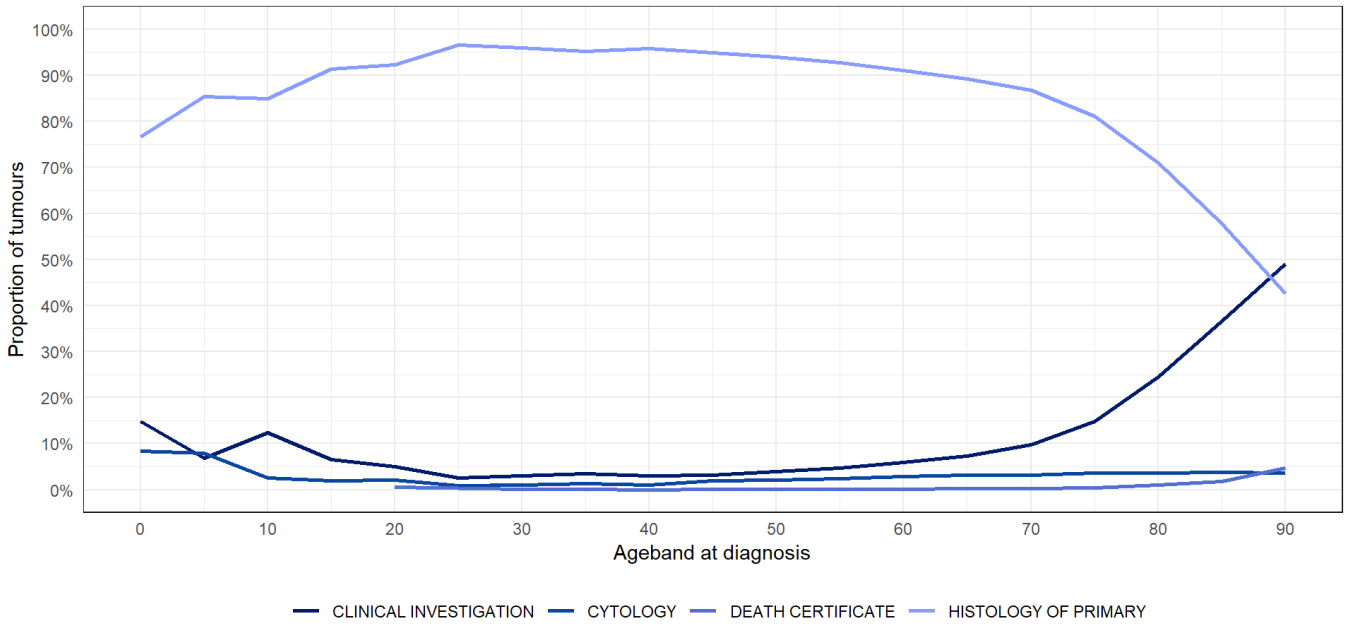
The marked reduction in the proportion of patients having their diagnosis confirmed from a pathological specimen (histology or cytology) explains the increase often observed at older ages in Figure A3, from the age of around 70, reflecting fewer patients having an invasive procedure performed on them 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: Public Health 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: Public Health England, National Cancer Registration and Analysis Service