Rapid Cancer Registration Dataset: data at 7th November 2020 (CAS2011)

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 18% of cases are missing and 5% 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 the four most 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 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

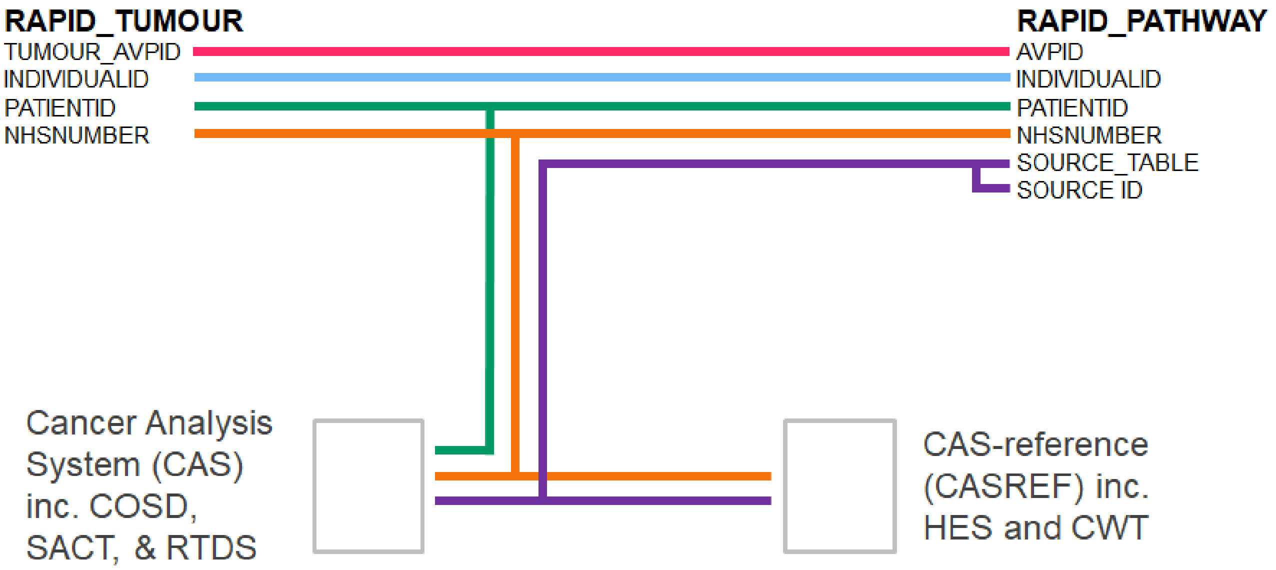
The rapid registration dataset consists of two tables:

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



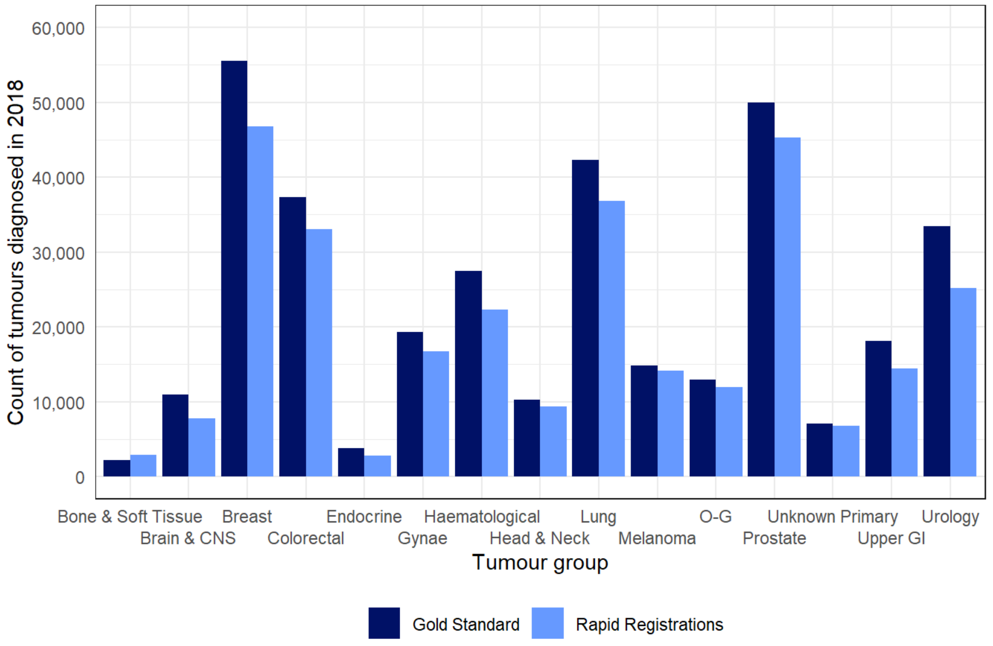
Data Quality

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. 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.

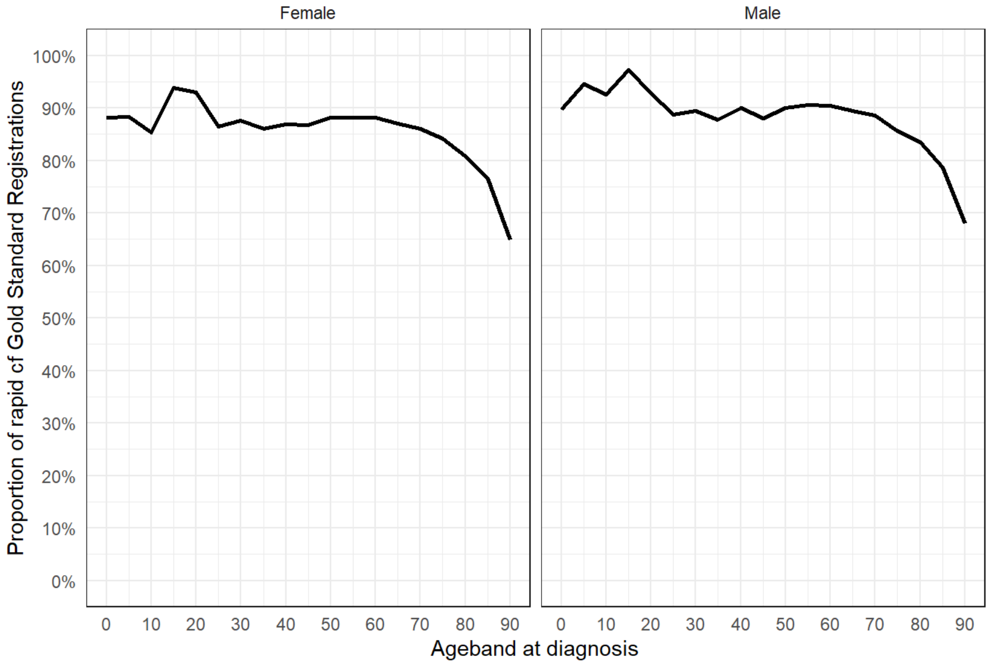
Figure 2: The number of cancer registrations by registration and tumour type, England, 2018



CNS: Central Nervous System; GI: Gastrointestinal; 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. 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

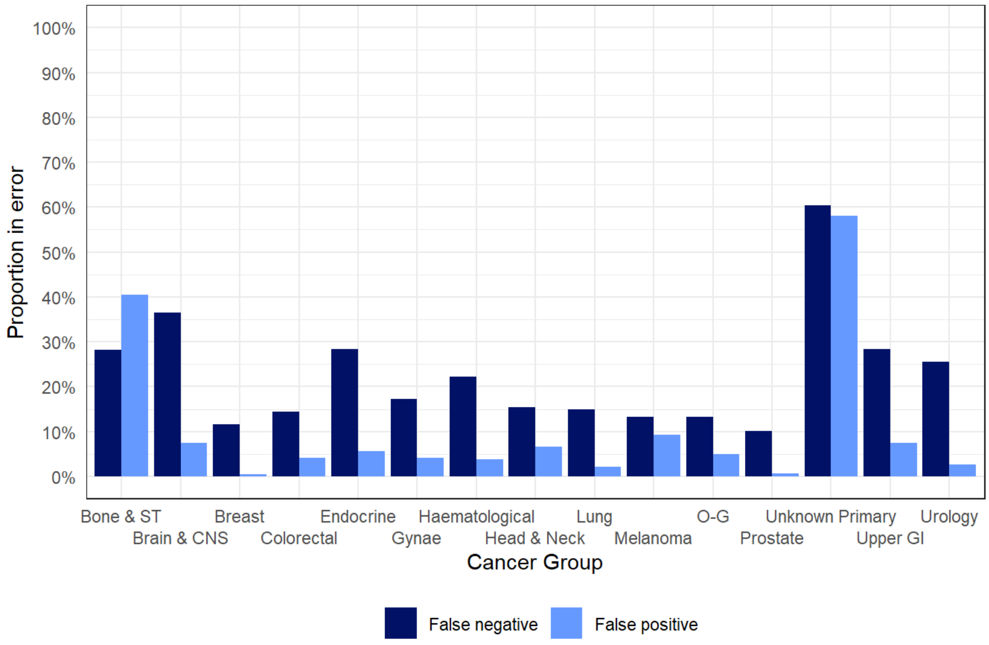
The quality of the Rapid Registrations was judged by comparing them against the gold-standard cancer registrations in the period April 2018 to September 2018. This period was chosen as available gold standard registration data was only finalised to December 2018 and a matching period of 90 days was allowed (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, measured in the cas2011 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

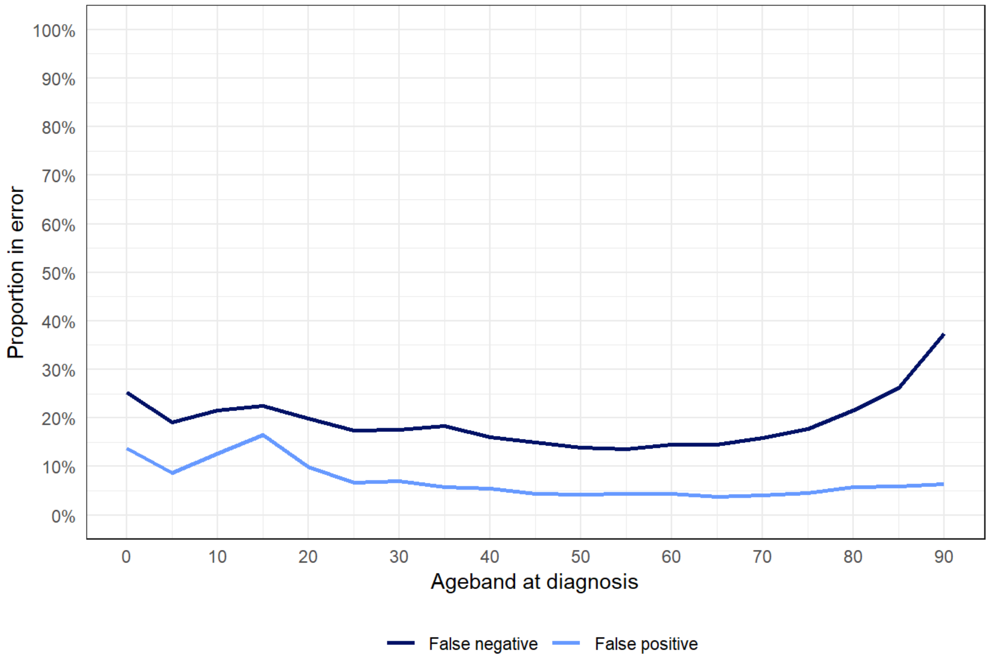


CNS: Central Nervous System; GI: Gastrointestinal; O-G: Oesophagogastric; ST: Soft Tissue
Source: Public Health England, National Cancer Registration and Analysis Service

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.

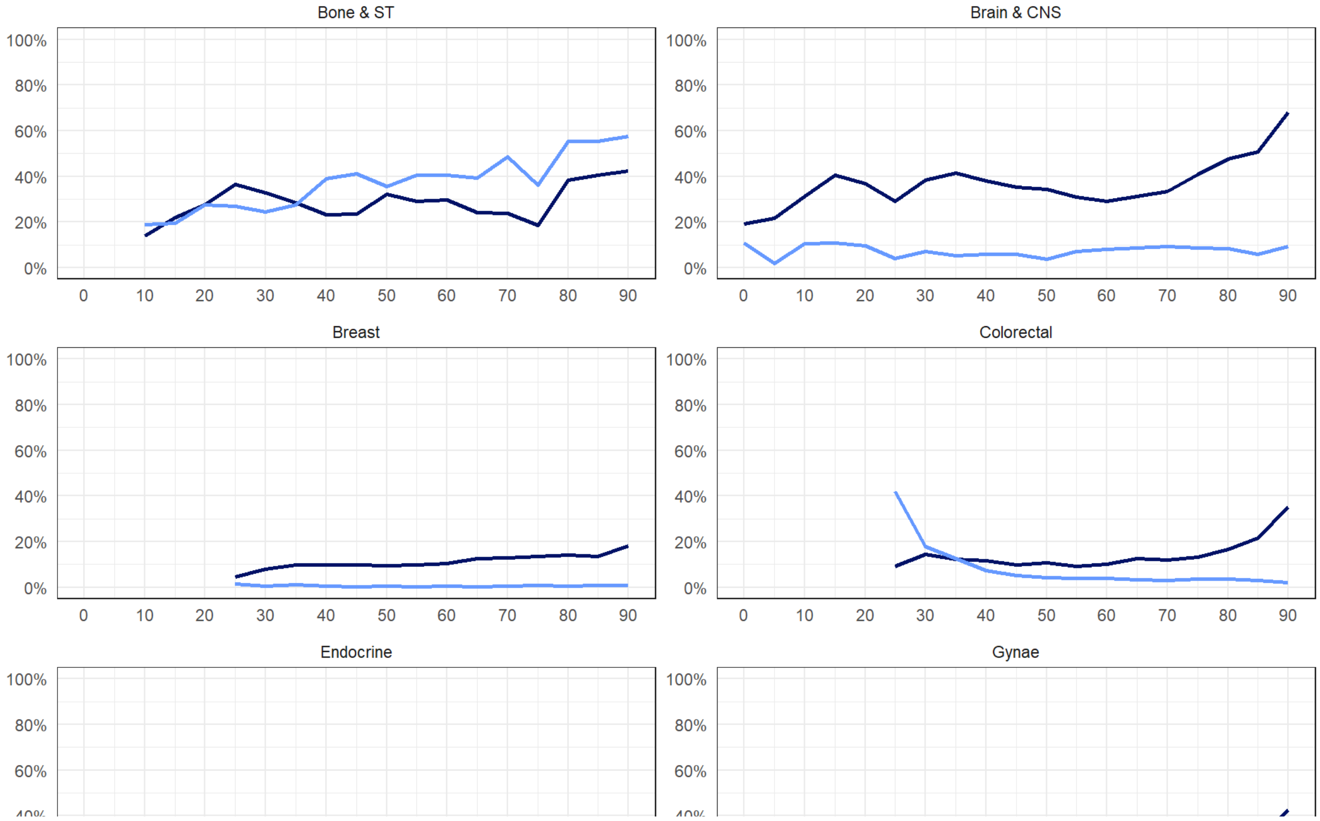
The patterns of false negative and false positive vary significantly by tumour group. Most groups have a higher proportion of false negatives than false positives at each age.

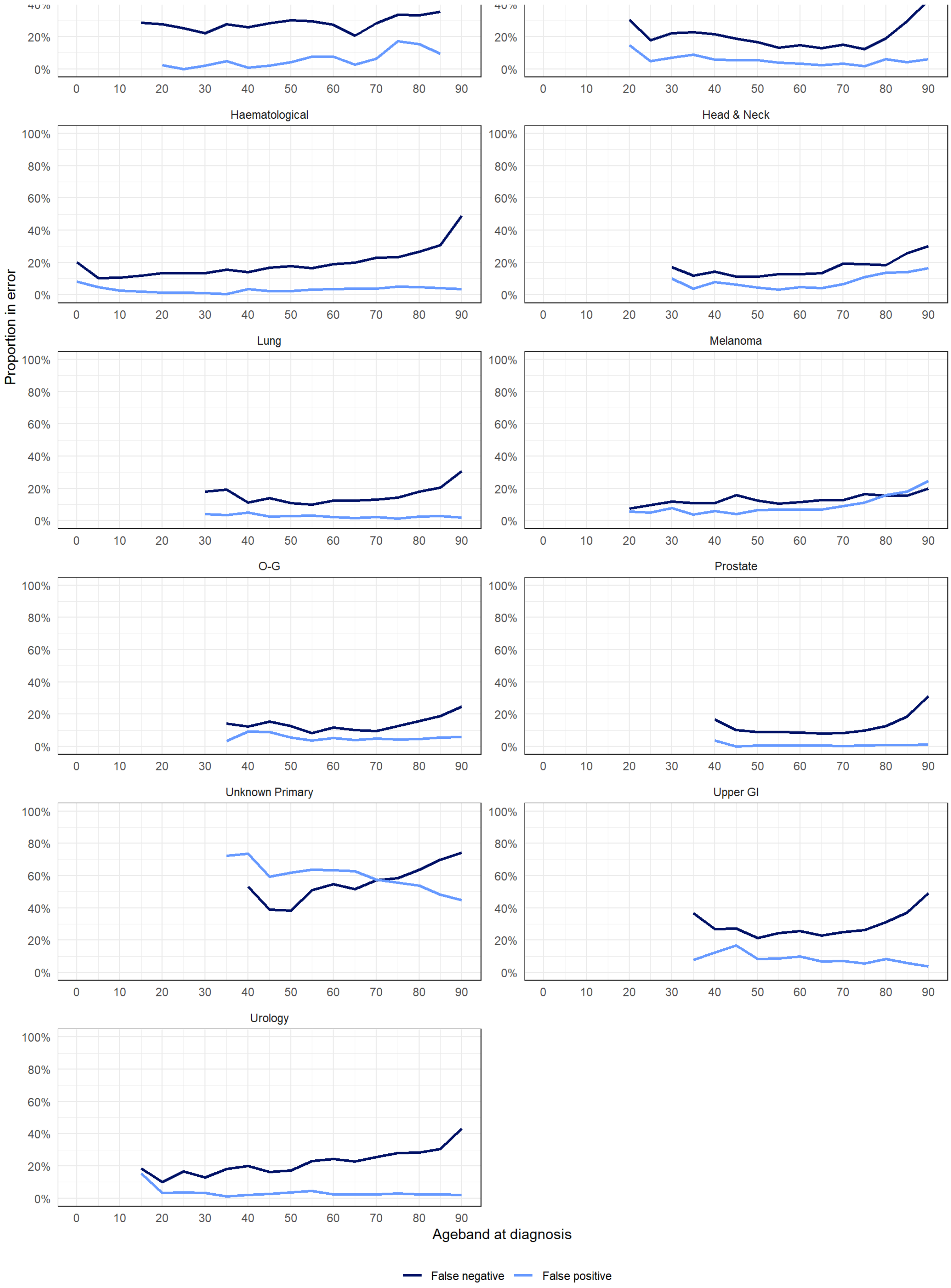
The proportion of false positives does not exhibit a trend by age for most tumour groups; the proportion rises with increasing age in the bone and soft tissue, head and neck groups and melanoma group and conversely falls with increasing age in the colorectal and unknown groups.

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

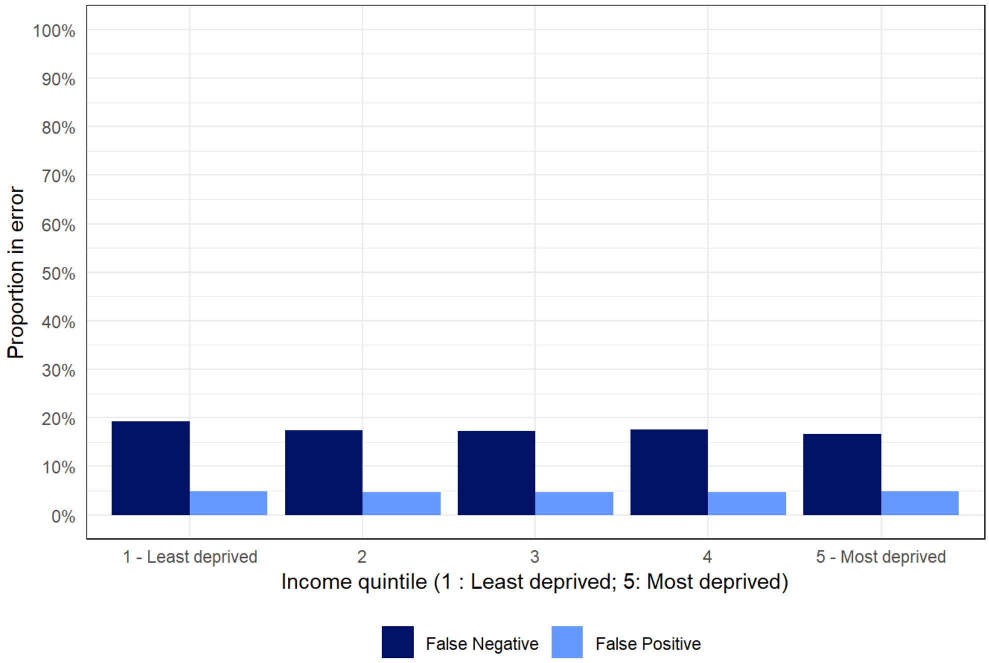




CNS: Central Nervous System; GI: Gastrointestinal; O-G: Oesophagogastric; ST: Soft Tissue
Source: Public Health England, National Cancer Registration and Analysis Service

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

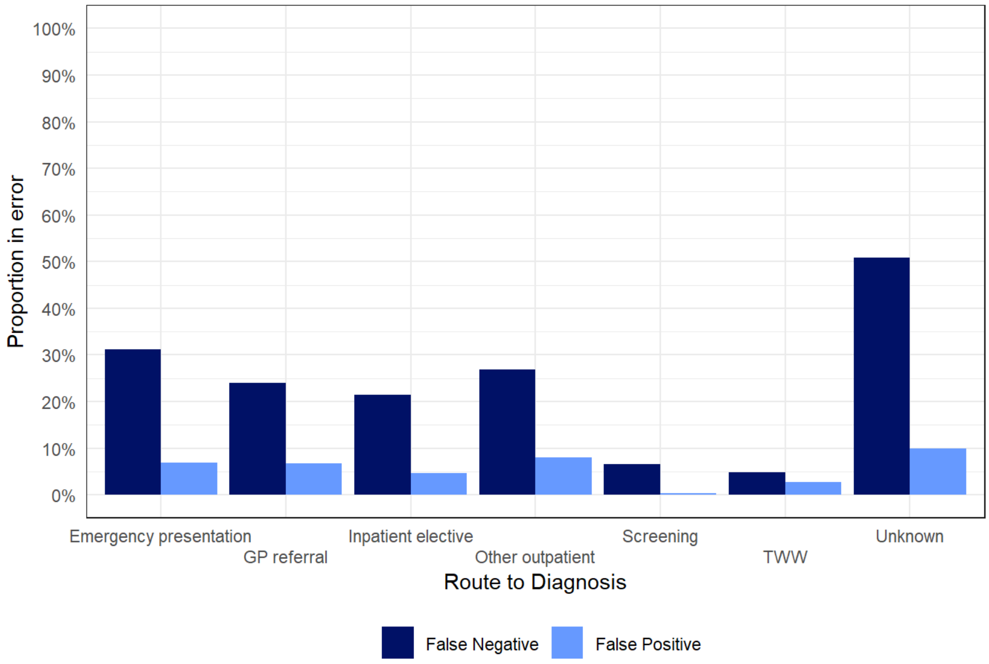


Source: Public Health England, National Cancer Registration and Analysis Service

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).

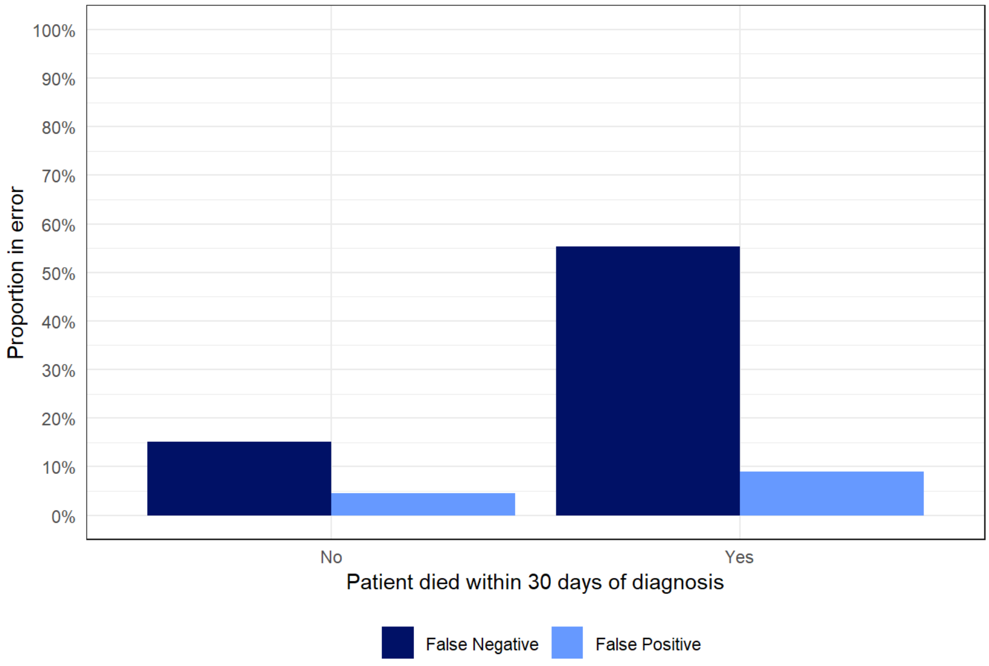
Figure 7: False negative and false positive errors by route to diagnosis



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.

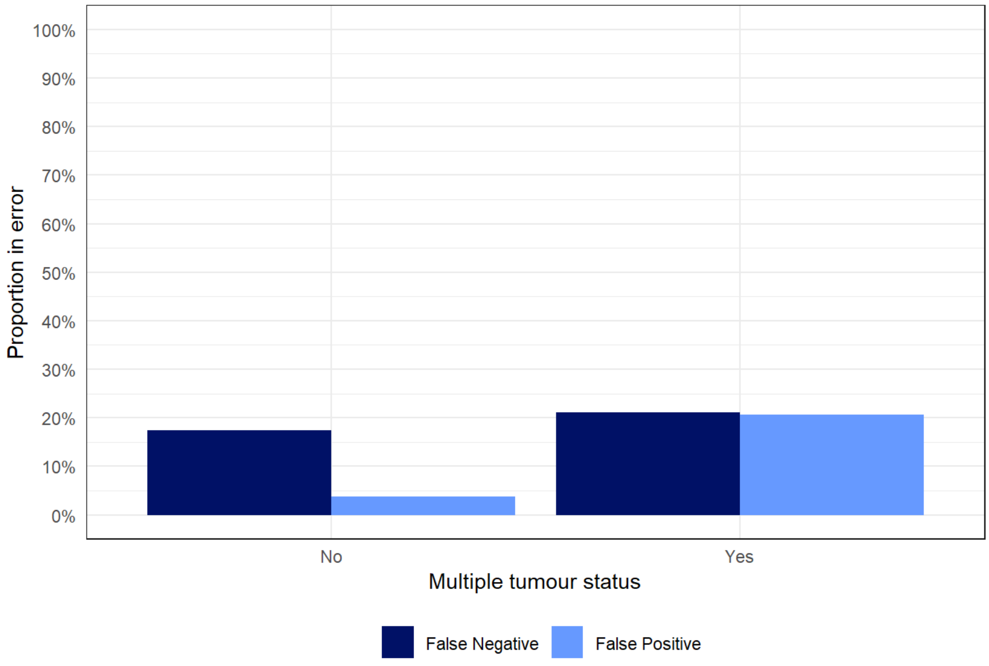
Figure 8: False negative and false positive errors by 30-day mortality



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.

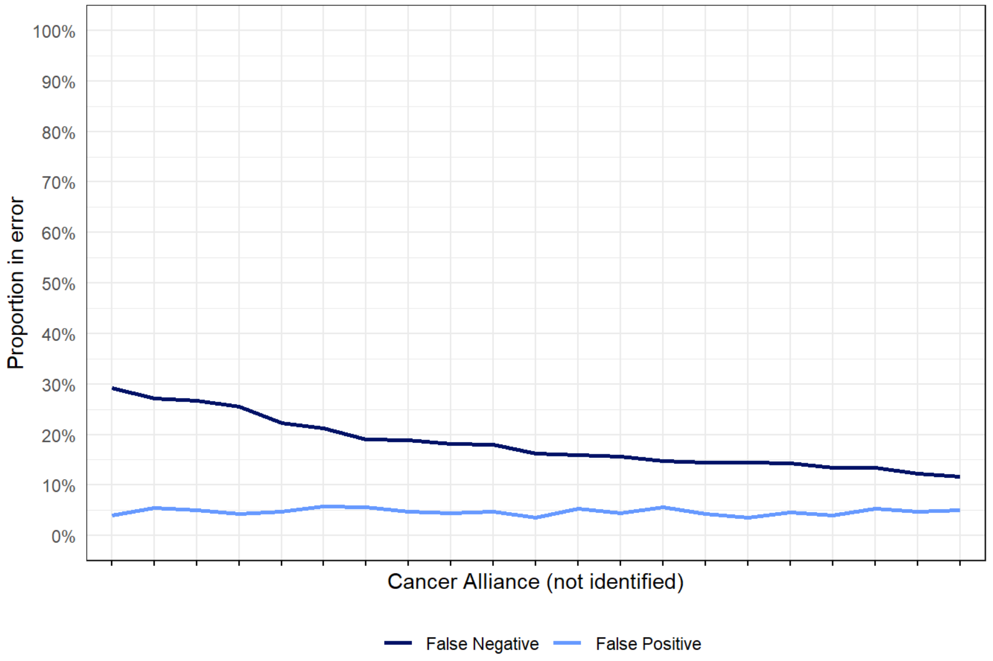
Figure 9: False negative and false positive errors by multiple tumour status



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

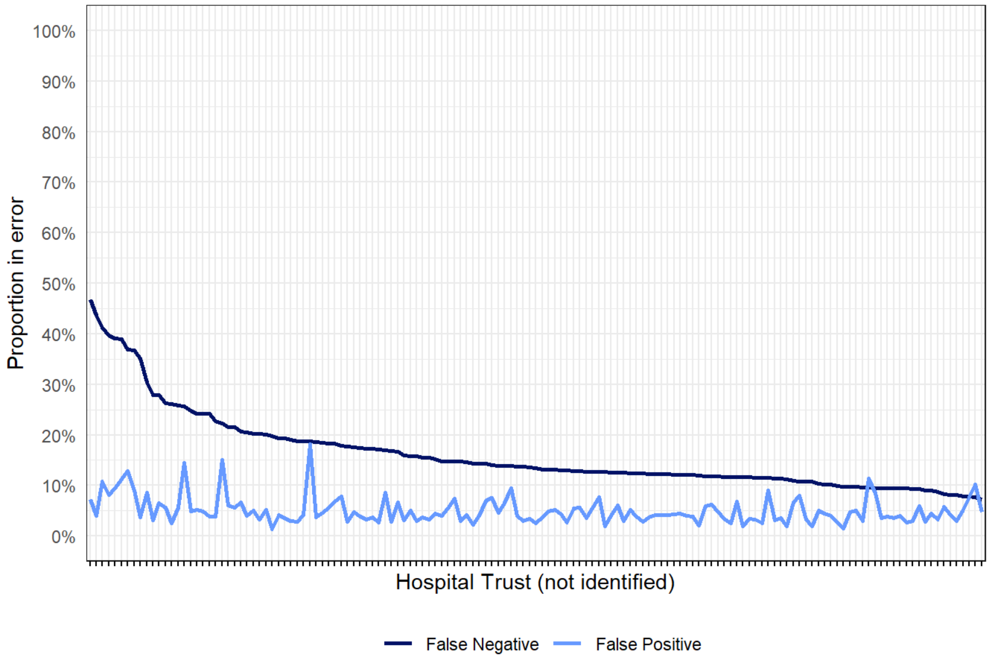


Source: Public Health England, National Cancer Registration and Analysis Service

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.

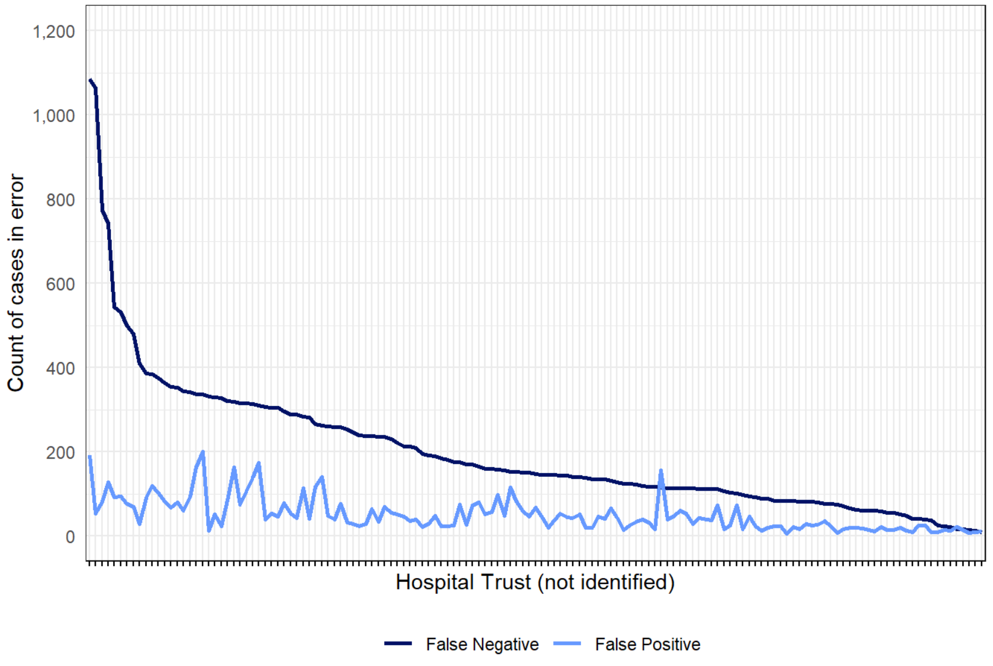
There is substantial variation in both false positive and false negative rates and counts. Some large trusts have several hundred or up to 1000 cases (over the six-month period under assessment).

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 %
Broad cancer group	90	17.7	5.2
None	90	16.2	3.1
Broad cancer group	60	19.0	6.6
Broad cancer group	30	23.8	11.9
Broad cancer group	14	33.9	23.4
Broad cancer group	7	49.5	41.6
Broad cancer group	0	83.1	80.2
3-digit ICD-10 code	90	25.0	13.0

Counts of events over time

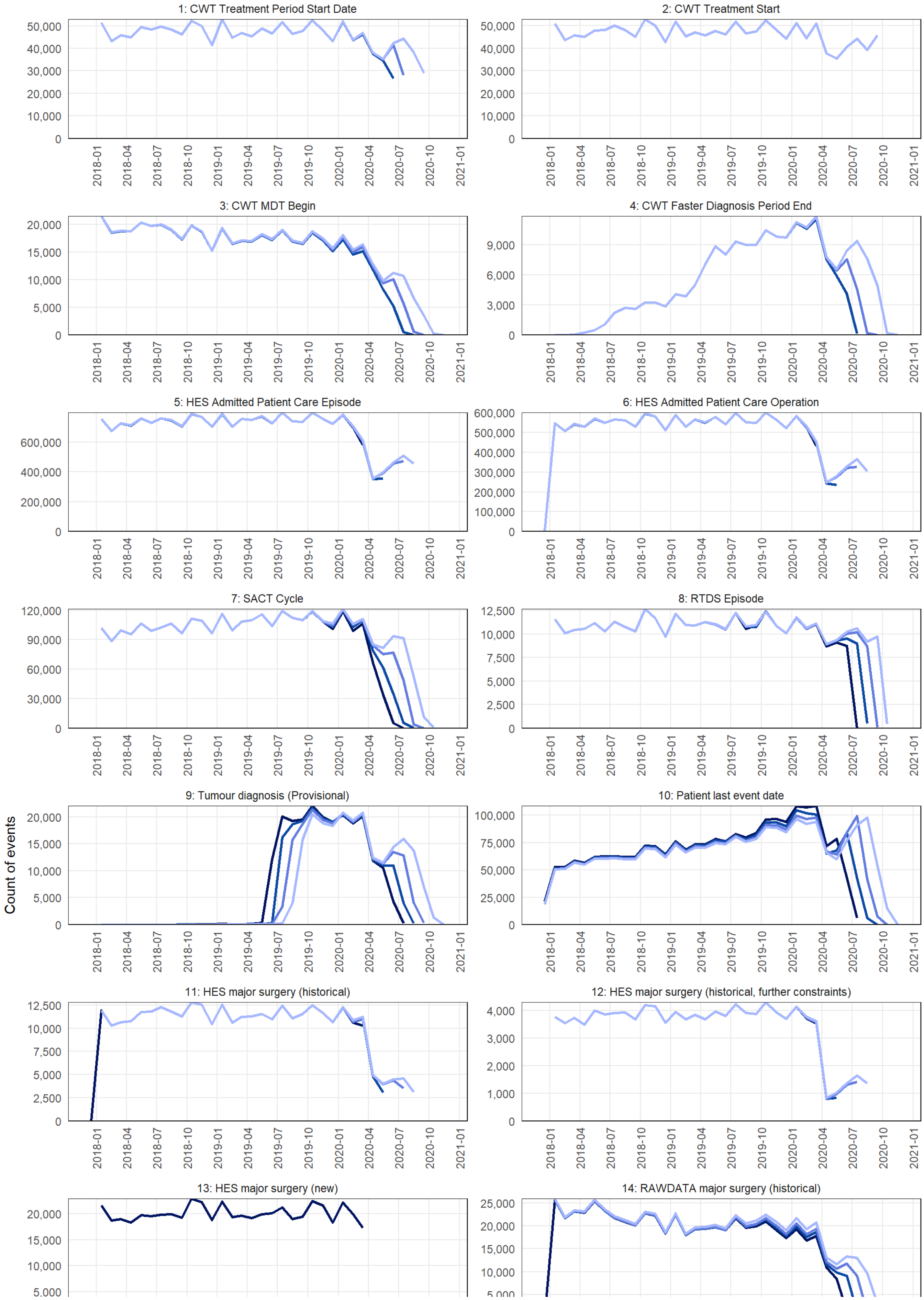
This section examines the population of events by chronological time and when they appear in successive analytical snapshots in the CAS. Figure 13 shows that most data items in the Rapid Registrations dataset are stable with respect to the snapshot month.

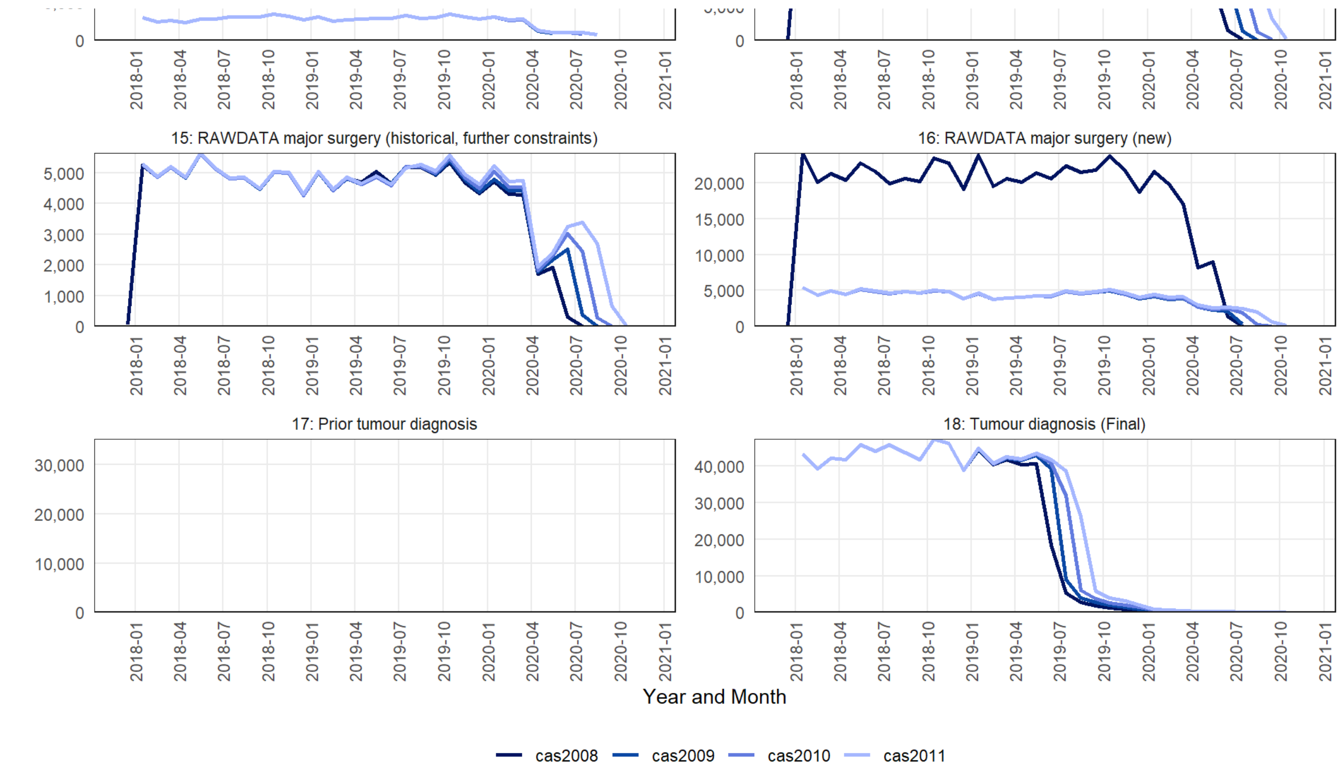
Specific comments about the events shown below are:

- 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.
- 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.

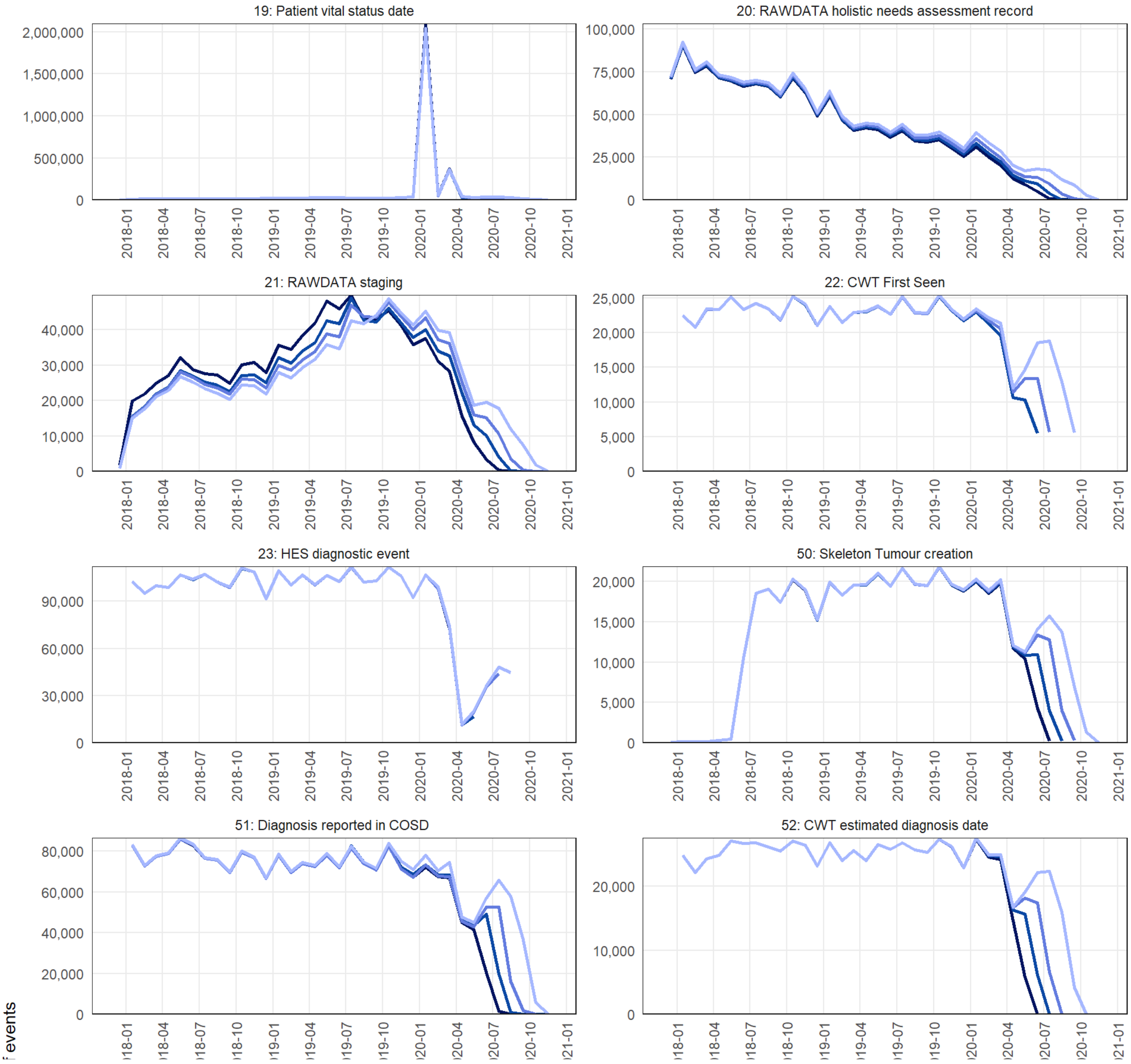
- The overall cohort was expanded from cas2009 to include a selection of D-codes, this is reflected in an increase in overall counts in (for example) Events 101-103.
- Operations on C44 tumours were removed from lookup tables generating events 13 and 16 from cas2010, this is reflected in large drop in event count overall.

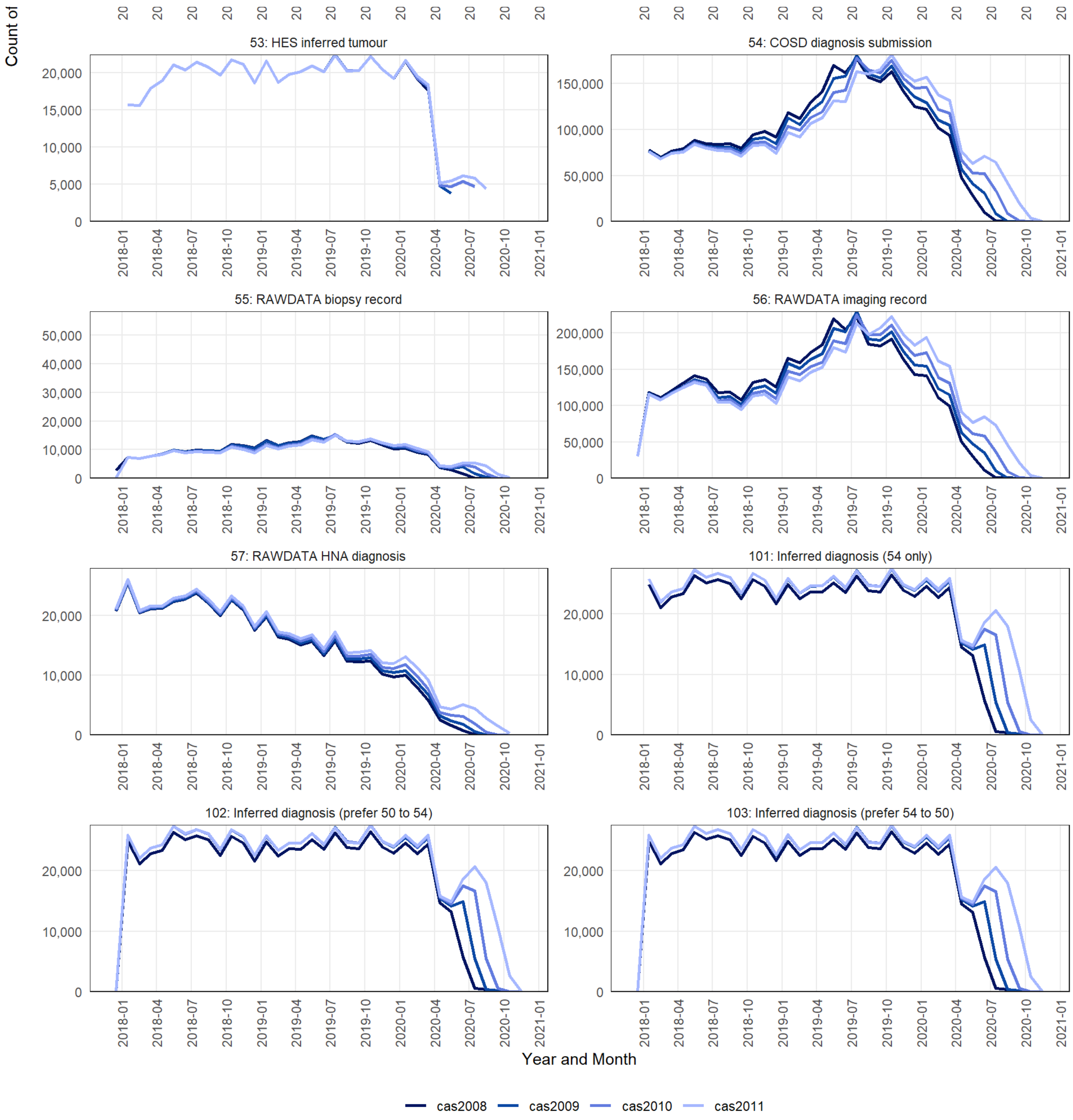
Figure 13: Population of data items to CAS snapshot





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, even in months prior to this though data is not 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 cas2011

Data source	Latest usable	April 2020	May 2020	June 2020	July 2020	August 2020	September 2020

Note:

TSD = Treatment Start Date

TPSD = Treatment Period Start Date

Data source	Latest usable	April 2020	May 2020	June 2020	July 2020	August 2020	September 2020
Rapid Tumours (COSD)	July 2020	Complete	Complete	Complete	98%	•	•
HES	July 2020	Complete	Complete	Complete	Complete	•	•
SACT	July 2020	94%	93%	90%	86%	•	•
RTDS	September 2020	Complete	Complete	98%	96%	96%	91%
CWT (TSD)	September 2020	Complete	Complete	Complete	Complete	Complete	Complete
CWT (TPSD)	August 2020	Complete	Complete	Complete	100%	98%	62%

Note:
TSD = Treatment Start Date
TPSD = Treatment Period Start Date

Table 3: Number of outlier CCGs in COSD dataset in cas2011

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
2019-07	1	0	134	135
2019-08	1	0	134	135
2019-09	0	1	134	135
2019-10	0	0	135	135
2019-11	0	0	135	135
2019-12	1	0	134	135
2020-01	0	0	135	135
2020-02	0	1	134	135
2020-03	0	1	134	135
2020-04	4	1	130	135
2020-05	1	1	133	135
2020-06	0	2	133	135
2020-07	0	3	132	135
2020-08	1	14	120	135
2020-09	44	49	42	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 breast, colorectal, lung and prostate cancer cases. This has been benchmarked against the gold standard cancer registry data for cas2011.

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)
Breast	1	6991	32.2%	8225	37.8%
Breast	2	6524	30.0%	8279	38.1%
Breast	3	1646	7.6%	1881	8.7%
Breast	4	546	2.5%	881	4.1%
Breast	U	6027	27.7%	2468	11.4%
Colorectal	1	2439	15.9%	2623	17.1%
Colorectal	2	3516	22.9%	3780	24.6%
Colorectal	3	4131	26.9%	4539	29.5%
Colorectal	4	2525	16.4%	3369	21.9%
Colorectal	U	2750	17.9%	1050	6.8%
Lung	1	3147	18.5%	3342	19.6%
Lung	2	1293	7.6%	1344	7.9%
Lung	3	3777	22.2%	3767	22.1%
Lung	4	7725	45.3%	8241	48.4%
Lung	U	1096	6.4%	344	2.0%
Prostate	1	6277	25.8%	8683	35.6%
Prostate	2	3098	12.7%	3618	14.8%
Prostate	3	5671	23.3%	6324	25.9%
Prostate	4	2851	11.7%	3978	16.3%
Prostate	U	6479	26.6%	1773	7.3%
All 4	1	18854	24.0%	22873	0.29
All 4	2	14431	18.4%	17021	0.22
All 4	3	15225	19.4%	16511	0.21
All 4	4	13647	17.4%	16469	0.21
All 4	U	16352	20.8%	5635	0.07

In Tables 5a-d 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 all four staged cancers except 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-d: Stage comparison between Rapid Registrations and Gold Standard Registrations by cancer site

a. breast

	Stage Group (Rapid)				
Stage Group (Gold Standard)	1	2	3	4	Unknown
1	89.3%	4.6%	1.3%	3.5%	27.3%
2	6.5%	88.9%	10.1%	14.3%	29.5%
3	0.6%	2.9%	81.4%	4.6%	4.8%
4	0.2%	0.8%	3.0%	71.6%	6.2%
U	3.5%	2.8%	4.2%	6.0%	32.2%

b. colorectal

	Stage Group (Rapid)				
Stage Group (Gold Standard)	1	2	3	4	Unknown
1	85.0%	1.8%	1.8%	0.7%	14.3%
2	5.5%	86.7%	5.7%	1.5%	11.8%
3	6.8%	6.8%	85.1%	4.1%	18.8%
4	0.8%	2.7%	5.4%	92.5%	25.3%
U	1.8%	2.0%	2.0%	1.3%	29.8%

c. lung

	Stage Group (Rapid)				
Stage Group (Gold Standard)	1	2	3	4	Unknown
1	93.5%	6.4%	1.0%	0.4%	22.4%
2	2.9%	85.6%	1.7%	0.4%	4.7%
3	1.6%	5.2%	90.5%	1.3%	12.0%
4	1.3%	2.2%	6.1%	97.5%	37.4%
U	0.8%	0.6%	0.7%	0.4%	23.5%

d. prostate

	Stage Group (Rapid)				
Stage Group (Gold Standard)	1	2	3	4	Unknown
1	86.8%	8.3%	3.8%	1.4%	42.0%
2	6.6%	84.2%	2.5%	0.9%	6.7%
3	4.1%	4.5%	87.4%	3.1%	13.6%
4	0.9%	0.7%	3.8%	92.7%	16.1%
U	1.6%	2.4%	2.5%	2.0%	21.6%

"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)
Breast	Early	13515	62.2%	16504	75.9%
Breast	Late	2192	10.1%	2762	12.7%
Breast	Unknown	6027	27.7%	2468	11.4%
Colorectal	Early	5955	38.8%	6403	41.7%
Colorectal	Late	6656	43.3%	7908	51.5%
Colorectal	Unknown	2750	17.9%	1050	6.8%
Lung	Early	4440	26.1%	4686	27.5%
Lung	Late	11502	67.5%	12008	70.5%
Lung	Unknown	1096	6.4%	344	2.0%
Prostate	Early	9375	38.5%	12301	50.5%
Prostate	Late	8522	35.0%	10302	42.3%
Prostate	Unknown	6479	26.6%	1773	7.3%
All 4	Early	33285	42.4%	39894	50.8%
All 4	Late	28872	36.8%	32980	42.0%
All 4	Unknown	16352	20.8%	5635	7.2%

Tables 7a-d: "Early" vs "late" stage comparison between Rapid Registrations and Gold Standard Registrations

a. breast

	Stage Category (Rapid)		
Stage Category (Gold Standard)	Early	Late	Unknown
Early	94.7%	13.0%	56.8%
Late	2.2%	82.4%	11.0%
Unknown	3.2%	4.7%	32.2%

b. colorectal

	Stage Category (Rapid)		
Stage Category (Gold Standard)	Early	Late	Unknown
Early	89.3%	5.5%	26.1%
Late	8.7%	92.8%	44.0%
Unknown	2.0%	1.7%	29.8%

c. lung

	Stage Category (Rapid)		
Stage Category (Gold Standard)	Early	Late	Unknown

	Stage Category (Rapid)		
Stage Category (Gold Standard)	Early	Late	Unknown
Early	95.1%	1.4%	27.1%
Late	4.1%	98.1%	49.4%
Unknown	0.7%	0.5%	23.5%

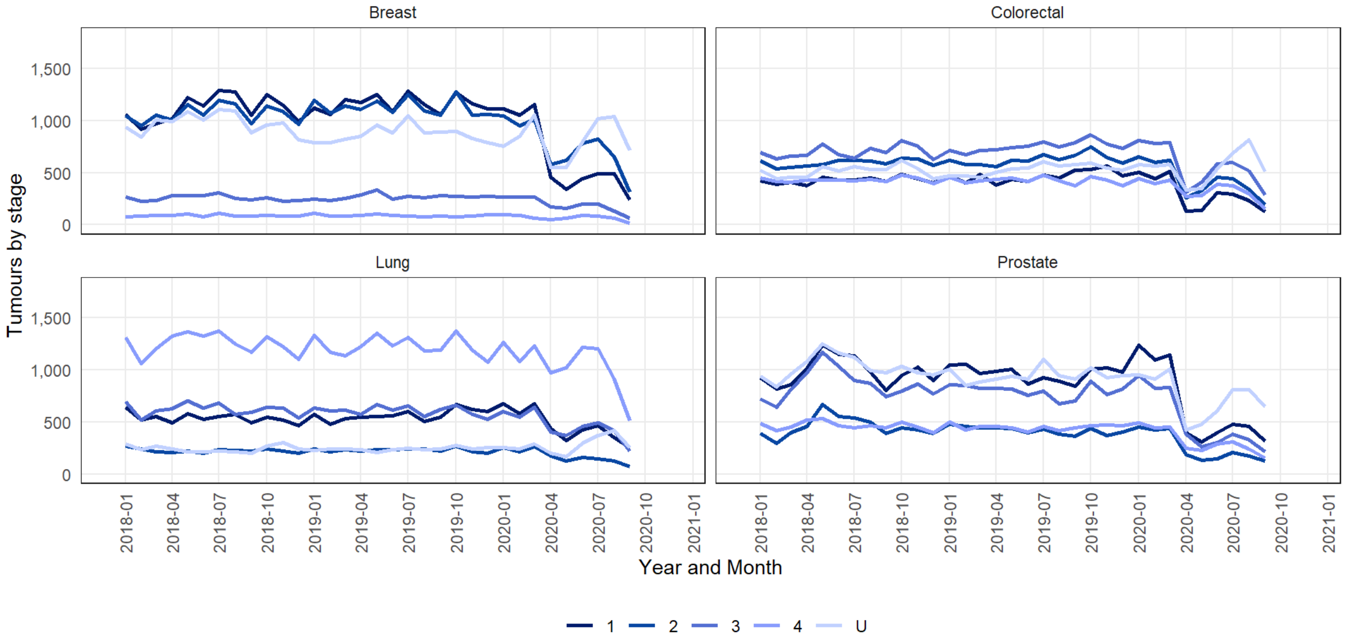
d. prostate

	Stage Category (Rapid)		
Stage Category (Gold Standard)	Early	Late	Unknown
Early	93.1%	4.9%	48.7%
Late	5.1%	92.7%	29.7%
Unknown	1.8%	2.3%	21.6%

Stage trends over time

Figure 13 shows the monthly variation of the incidence count by stage at diagnosis for the four most common cancers (excluding non-melanoma skin cancer). 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



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

EVENT_TYPE	EVENT_DESC	EVENT_PROPERTY_1	EVENT_PROPERTY_2	EVENT_PROPERTY_3	EVENT_DATE	Linkage
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			Vitalstatusdate	PATIENTID
20	RAWDATA holistic needs assessment record	HNA point of pathway **	Primary diagnosis	Laterality	Date of HNA	PATIENTID

EVENT_TYPE	EVENT_DESC	EVENT_PROPERTY_1	EVENT_PROPERTY_2	EVENT_PROPERTY_3	EVENT_DATE	Linkage
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

COLUMN_NAME	DATA_TYPE	Notes
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 malignant breast, colorectal, lung and prostate cancer
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
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

ICD	CANCER_GROUP	ICD	CANCER_GROUP
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
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

ICD	CANCER_GROUP	ICD	CANCER_GROUP
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.

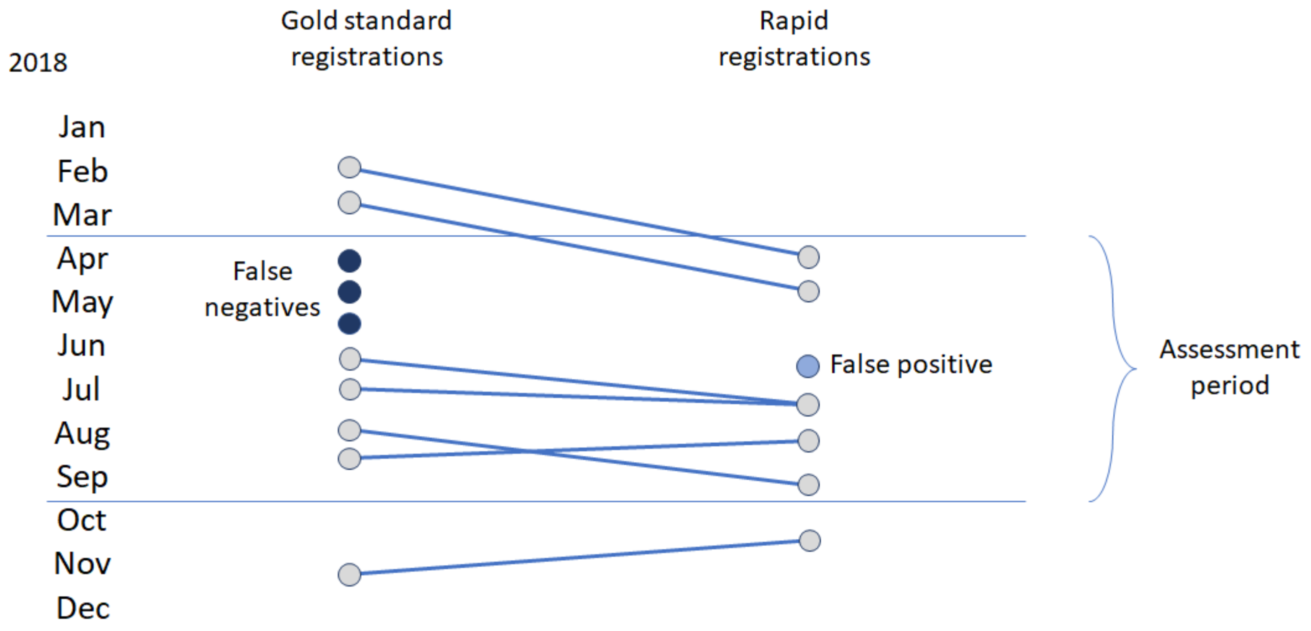
Table A4: Rapid Cancer Registrations: alternative defining events

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 - filtered COSD	5.2%	17.6%

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	5362	3764	1598	70.2%	378	1959
Breast	28863	24284	4579	84.1%	213	3348
Colorectal	18849	16535	2314	87.7%	743	2728
Endocrine	1885	1394	491	74.0%	105	536
Gynae	9723	8308	1415	85.4%	395	1680
Haematological	13642	11072	2570	81.2%	450	3037
Head & Neck	5255	4717	538	89.8%	329	813
Lung	21449	18584	2865	86.6%	479	3214
Melanoma	8099	7555	544	93.3%	728	1079
O-G	6600	5991	609	90.8%	311	878
Prostate	26785	24341	2444	90.9%	176	2699
Bone & Soft Tissue	1133	1350	-217	119.2%	555	320
Unknown Primary	3608	3357	251	93.0%	1957	2178
Upper GI	9137	7161	1976	78.4%	596	2598
Urology	16809	12664	4145	75.3%	467	4304

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	140	-31	128.4%	56	24
C01	641	438	203	68.3%	9	89
C02	603	604	-1	100.2%	16	91
C03	232	104	128	44.8%	5	70
C04	250	236	14	94.4%	11	35
C05	214	180	34	84.1%	7	36
C06	267	278	-11	104.1%	18	52
C07	236	261	-25	110.6%	75	54
C08	81	84	-3	103.7%	13	14
C09	910	731	179	80.3%	13	92
C10	150	226	-76	150.7%	10	37
C11	110	100	10	90.9%	3	18
C12	154	98	56	63.6%	1	15
C13	143	123	20	86.0%	10	30
C14	24	57	-33	237.5%	11	14
C15	3989	4020	-31	100.8%	102	416
C16	2611	1971	640	75.5%	209	462
C17	799	627	172	78.5%	121	279
C18	12352	10849	1503	87.8%	559	1997

Cancer site	Gold Standard (GS) Registrations	Rapid Registrations	Difference	Percentage Rapid/GS	FPE	FNE
C19	987	800	187	81.1%	19	160
C20	4866	4276	590	87.9%	88	508
C21	644	610	34	94.7%	77	63
C22	2589	2037	552	78.7%	219	812
C23	473	408	65	86.3%	27	114
C24	640	470	170	73.4%	27	142
C25	4486	3488	998	77.8%	108	1130
C26	150	131	19	87.3%	94	121
C30	161	145	16	90.1%	21	30
C31	92	59	33	64.1%	4	28
C32	878	853	25	97.2%	46	84
C33	13	10	3	76.9%	1	3
C34	20001	17316	2685	86.6%	426	2953
C37	166	82	84	49.4%	9	61
C38	74	327	-253	441.9%	31	36
C39	NA	13	NA	NA%	4	NA
C40	118	104	14	88.1%	11	25
C41	114	181	-67	158.8%	113	41
C43	8099	7555	544	93.3%	728	1079
C45	1195	836	359	70.0%	8	161
C46	68	45	23	66.2%	4	26
C47	25	14	11	56.0%	6	19
C48	283	366	-83	129.3%	102	96
C49	833	1020	-187	122.4%	427	228
C50	25050	21770	3280	86.9%	184	2736
C51	639	491	148	76.8%	23	148
C52	93	91	2	97.8%	9	20
C53	1302	1171	131	89.9%	34	188
C54	4093	3508	585	85.7%	72	366
C55	74	291	-217	393.2%	16	32
C56	2965	2087	878	70.4%	101	769
C57	264	281	-17	106.4%	21	58
C58	10	22	-12	220.0%	17	3
C60	302	278	24	92.1%	31	56
C61	26785	24341	2444	90.9%	176	2699
C62	1052	996	56	94.7%	62	112

Cancer site	Gold Standard (GS) Registrations	Rapid Registrations	Difference	Percentage Rapid/GS	FPE	FNE
C63	29	16	13	55.2%	6	24
C64	4755	3884	871	81.7%	191	1038
C65	403	293	110	72.7%	17	110
C66	353	227	126	64.3%	9	139
C67	4438	4657	-219	104.9%	93	975
C68	93	46	47	49.5%	3	47
C69	367	326	41	88.8%	34	60
C70	20	36	-16	180.0%	6	8
C71	2240	1788	452	79.8%	154	577
C72	76	71	5	93.4%	27	23
C73	1720	1299	421	75.5%	62	436
C74	113	59	54	52.2%	19	69
C75	52	36	16	69.2%	24	31
C76	94	524	-430	557.4%	429	76
C77	300	334	-34	111.3%	238	95
C78	679	213	466	31.4%	164	462
C79	286	331	-45	115.7%	238	202
C80	2249	1955	294	86.9%	888	1343
C81	895	824	71	92.1%	6	100
C82	1198	1006	192	84.0%	6	166
C83	3140	2556	584	81.4%	26	495
C84	382	211	171	55.2%	10	139
C85	1338	781	557	58.4%	32	440
C86	NA	90	NA	NA%	3	NA
C88	194	353	-159	182.0%	8	43
C90	2496	1929	567	77.3%	29	615
C91	2129	1694	435	79.6%	42	487
C92	1734	1205	529	69.5%	75	492
C93	23	142	-119	617.4%	7	5
C94	26	122	-96	469.2%	104	9
C95	50	35	15	70.0%	1	27
C96	37	124	-87	335.1%	101	19
D05	3813	2514	1299	65.9%	29	612
D09	4879	407	4472	8.3%	33	1565
D32	1292	702	590	54.3%	31	591
D33	401	471	-70	117.5%	59	188

Cancer site	Gold Standard (GS) Registrations	Rapid Registrations	Difference	Percentage Rapid/GS	FPE	FNE
D35	441	250	191	56.7%	29	226
D41	505	1860	-1355	368.3%	22	238
D42	134	6	128	4.5%	1	54
D43	260	77	183	29.6%	18	139
D44	106	23	83	21.7%	13	74

Appendix 6 - False negative errors and basis of diagnosis

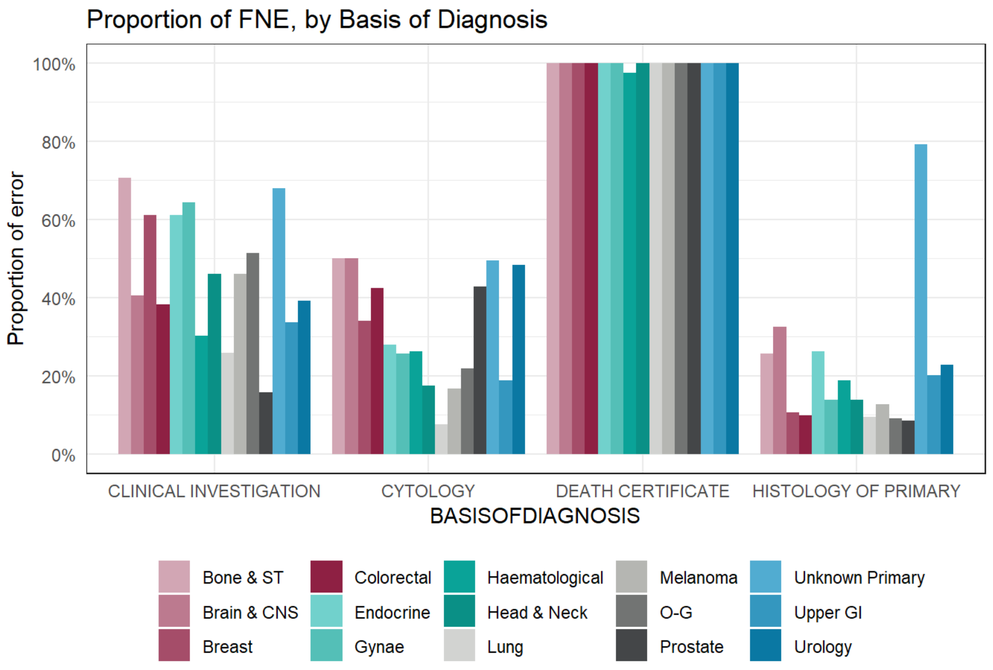
This appendix explores the reason for the overall age-dependence of the false negative error rate.

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.

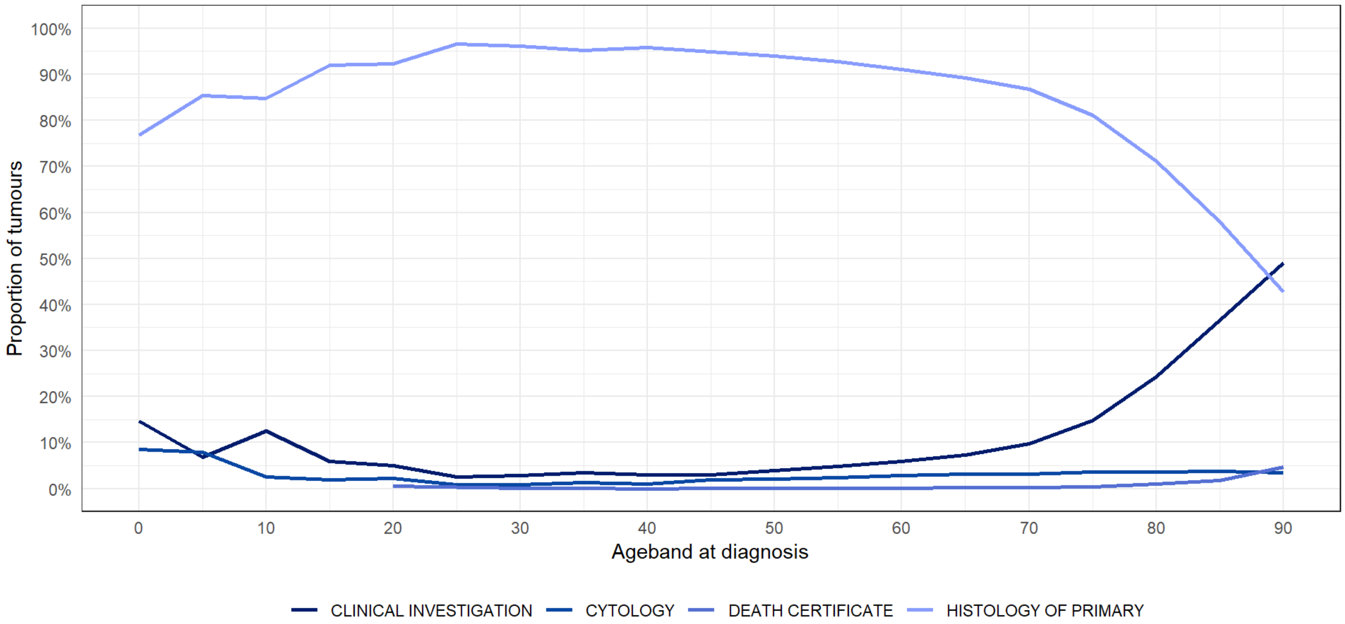
The marked reduction in the proportion of patients having their diagnosis confirmed from a pathological specimen (histology or cytology) explains the increase often observed at older ages in Figure A3, from the age of around 70, reflecting fewer patients having an invasive procedure performed on them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