The completeness, concordance, and timeliness of cancer diagnosis data collected on-site during a prospective cohort study compared to central cancer registries in England and Wales

Ashley Jackson

St John's College
University of Oxford

A thesis submitted to the Division of Medical Sciences at the University of Oxford for the degree of Master of Science by Research in Oncology

Trinity 2024

Dedication

This thesis is dedicated to my dad—whose battle with cancer was my inspiration to study oncology and continues to be a driving force in my life.

Acknowledgements

First and foremost, I would like to thank my supervisors, Dr. Brian Nicholson and Dr. Eileen Parkes, for their guidance and support throughout this project. I am grateful to them for helping me find a project that best suited my interests and goals, for giving me many opportunities to grow as a researcher, and for supporting me in presenting my work to other members of the scientific community. It has been a pleasure to conduct my master's work under their guidance.

I would like to give a special thanks to Dr. Pradeep Virdee and Dr. Sharon Tonner for their continued guidance and technical support throughout the course of this work. Pradeep provided invaluable assistance with my analyses, while Sharon was a constant support, liaising with the many collaborators on this project. This thesis work would not have been possible without their contributions.

I would like to acknowledge our collaborators on this project, GRAIL Bio UK and the cancer registries of England and Wales, who provided helpful feedback on my manuscript and valuable insight into the cancer registration process. I would also like to thank all collaborators and authors on the initial SYMPLIFY study, as well as the patients who were involved in the study, as they provided the groundwork for my thesis work.

This work would not have been possible without the funding I received from the Rhodes Trust, for which I am extremely grateful.

Finally, I would like to thank my incredible family, friends, and co-workers who supported me throughout this work and made it such an enjoyable process. I would like to thank my mom, for I wouldn't have made it to Oxford in the first place without her unrelenting support all my life. Lastly, I would like to thank my partner, Jeremy, who has been my rock throughout this entire process. If it weren't for him agreeing to pick up and move from Ottawa to Oxford with me, this degree wouldn't have been possible.

Table of Contents

List of Abbreviations

2WW	Two-week-wait
ATBC	Alpha-Tocopherol Beta-Carotene Cancer Prevention
CAP	Cluster randomised triAl of prostate specific antigen testing for Prostate cancer
CDC	Centres for Disease Control and Prevention
CNS	Central nervous system
CONSORT	Consolidated Standards of Reporting Trials
COSD	Cancer Outcomes and Services Dataset
CRF	Case report form
CRN	Cancer Registry of Norway
CRO	Cancer registration officer
CT	Computed tomography
CWT	Cancer Waiting Times
DCR	Danish Cancer Registry
DHCW	Digital Health and Care Wales
ENETS	European Neuroendocrine Tumour Society
GP	General practitioner
FCR	Finnish Cancer Registry
FIGO	International Federation of Gynaecology and Obstetrics
HES	Hospital Episodes Statistics
ICD	International Classification of Diseases
ICD-10	International Classification of Diseases, 10th Revision
ICD-O-3	International Classification of Diseases for Oncology, 3rd Edition
ISS	International Staging System
MCED	Multi-cancer early detection
NBCR	National Breast Cancer Register
NBH	National Board of Health
NCDB	National Cancer Database
NCRAS	National Cancer Registration and Analysis Service
NCRD	National Cancer Registration Dataset

NHS	National Health Service
NHSCR	National Health Service Central Register
NICRD	Northern Ireland Cancer Registry
NPOCR	National Program of Cancer Registries
NPCR	National Prostate Cancer Register
NREV	National Register for Oesophageal and Gastric Cancer
NZCCR	New Zealand Children's Cancer Registry
NZCR	New Zealand Cancer Registry
PET	Positron emission tomography
RCRD	Rapid Cancer Registration Dataset
ROC	Regional Oncological Centres
ROUTINE	Randomised Controlled Trials Conducted Using Routinely Collected Data
SCRCR	Swedish Colorectal Cancer Registry
SEEGCR	Spanish EURECCA Esophagogastric Cancer Registry
SRCR	Swedish Rectal Cancer Registry
SWEDEHEART	Swedish web-system for the enhancement and development of evidence-based care in heart disease according to recommended therapies
TASTE	Thrombus Aspiration during ST-Segment Elevation Myocardial Infarction
TNM	Tumour, nodes, metastasis
UICC	Union for International Cancer Control
UKCTOCS	UK Collaborative Trial of Ovarian Cancer Screening
UKIACR	United Kingdom and Ireland Association of Cancer Registries
WBCR	Waikato Breast Cancer Register
WB-MRI	Whole-body magnetic resonance imaging
WCISU	Welsh Cancer Intelligence and Surveillance
WHO	World Health Organisation

List of Figures

List of Tables

Abstract

The completeness, concordance, and timeliness of cancer diagnosis data collected on-site during a prospective cohort study compared to central cancer registries in England and Wales

Submitted for the degree of Master of Science by Research in Oncology

Ashley Jackson, St John's College, Trinity 2024

Cancer places a high burden on healthcare systems and impacted individuals, leading to significant resources being committed to cancer research. Using central databases, such as national cancer registries, is one possible method for improving the resource burden associated with clinical cancer research. I compared the completeness, concordance, and timeliness of diagnostic cancer data obtained on-site to the national registries of England and Wales to determine when it is most appropriate to use these resources. On-site data were collected from 44 hospital sites during a prospective cohort study in England and Wales (SYMPLIFY; ISRCTN10226380). Linked central data was obtained from Digital Health and Care Wales (DHCW), the Welsh Cancer Intelligence and Surveillance Unit (WCISU), and the English National Cancer Registration Dataset (NCRD) and Rapid Cancer Registration Dataset (RCRD). Data cuts from these datasets were retrieved regularly between April 2022 and September 2023. Four data fields were investigated if recorded: International Classification of Diseases, 10th Revision (ICD-10) code, International Classification of Diseases for Oncology, 3rd Edition (ICD-O-3) morphology code, cancer stage, and Tumour, Nodes, and Metastasis (TNM) classification. ICD-10 completeness was 100% for each dataset, while completeness of the other variables varied between datasets. Concordance was highest between SYMPLIFY and registry data for ICD-10 code and broad morphology groupings but was less so for specific ICD-O-3 code, cancer stage, and TNM classification. The SYMPLIFY dataset reached full completion at 12 months post-enrolment, compared to 13 months for NCRD. I report comparable completeness and timeliness between on-site data collected from the SYMPLIFY study and national cancer registries, as well as high concordance for ICD-10 code and ICD-O-3 morphology groupings. These findings suggest that diagnostic cancer data from central registries in England and Wales may be able to support clinical research and help alleviate the resource burden associated with cancer research.

1. Introduction

Cancer is one of the leading causes of death worldwide. It is estimated that half of all people in the United Kingdom will get cancer at some point in their lifetime, and around 10 million people worldwide die from cancer every year (1, 2). Given the significant burden that cancer places on society and the people affected, there is a need to develop newer and more effective treatments and diagnostic modalities to improve the survival and quality of life for cancer patients. As such, a considerable amount of resources and time is dedicated to cancer research.

The economic burden of cancer research is substantial. Between 2016 and 2020, an estimated $24.5 billion was spent on cancer research globally (3). In 2022/2023, the world's largest independent cancer research charity, Cancer Research UK, spent £415 million on cancer research (4). This funding goes towards a variety of research activities, including pre-clinical research, clinical trials, public health research, and more, and spans various cancers and treatment types (3). Clinical trials are a vital part of testing new cancer therapeutics for their safety and efficacy. A prior study found that it costs more than $6,000 per patient enrolled in a clinical trial, and nearly a third of this is devoted to non-clinical costs (5).

Cancer research is not only costly but also a time-consuming process. Studies suggest that approximately 200 hours of work is associated with each patient enrolled in a clinical trial (5). Interestingly, nearly a third of this time is dedicated to non-clinical activities, such as filling out case reporting forms, data cleaning, and data management (5). These statistics illustrate the need to improve the efficiency and cost-effectiveness of cancer research. One method that has been suggested to improve the resource burden associated with cancer research is using central databases, such as national cancer registries.

1.1. Cancer Registries and Registry-Based Research

Cancer registries are massive repositories that include information on most, if not all, cancers diagnosed among individuals residing in a given area. Registries collect information on all of the people diagnosed with cancer in an area, including personal details such as name and sex, the cancer diagnosis, such as the cancer type and stage, cancer treatments, such as the types and durations of treatments applied, and the outcomes of the patients, such as details of remission or death (6). Although the acquisition of this data varies based on the given registry, information is primarily obtained from medical records and healthcare professionals. Medical registries play an essential role in public knowledge about diseases. Cancer registries allow for the monitoring and surveillance of cancer within a population and help to guide funding decisions and public health measures, making them extremely useful and comprehensive cancer databases (7).

Given that cancer registries include detailed information on all the cancers diagnosed in a given area, they represent an invaluable resource in cancer research. Potential uses of these national databases include the implementation of registry-based trials. Registry-based trials are those that use existing national medical registries for the recruitment, data collection, and follow-up of study participants (8). Registries allow a platform for carrying out several types of research studies, including cluster, crossover, parallel arm, and stepped-wedge trials (9). Given the growing interest in registry-based trials, an extension on the Consolidated Standards of Reporting Trials (CONSORT) extension for randomised controlled trials conducted using routinely collected data (CONSORT-ROUTINE) was developed (10).

One study in the cardiology setting modelled the potential cost-savings of using registry-based trials compared to standard trials and found that the savings varied anywhere from

approximately $4,300 to $600,000 depending on the study size and characteristics and were more cost-effective than standard trials 98.6% of the time (11). The models showed that factors that further enhanced the cost-effectiveness of registry-based trials included an increase in the number of patients in the study, an increase in the number of data fields to be extracted, and an increase in the amount of time it would take researchers to extract the data points manually (11). Other studies have also found registry-based randomised controlled trials to be more cost-effective than standard trials, and some of the reasons for these cost-savings included a reduced need for follow-up visits and monitoring, due to outcomes information being obtained from the registry, and limited staff training for data collection (12-14). One such study, the Thrombus Aspiration during ST-Segment Elevation Myocardial Infarction (TASTE) study, used a registry-based trial design, enrolling patients and evaluating endpoints through national registries. The cost of the study was very low at around $50 per patient—much lower than the previously referenced estimate of about $6,000 per patient per standard clinical trial (8, 15). Meanwhile, another trial using routinely-collected health data, investigating targeted versus universal decolonisation to prevent hospital infections, reduced their study costs to about $40 per patient through the cost-effective study design (16). Clearly, registry-based trials offer an economical way to conduct high-impact research.

Although low cost is one of the most apparent advantages of registry-based research, there are other benefits to using registry data for trials. Given their vast information on an entire population, the use of registries in research allows for a more inclusive study design that enhances the generalisability and external validity of the study findings (8, 17). The more diverse and inclusive study populations that registry-based trials may afford may help combat biases and issues of lack of diversity in current clinical trials and ultimately improve the quality and impact

of the research produced (18, 19). In addition to fostering a more inclusive and representative study, using medical registries in clinical study design can allow for more rapid study enrolment. Given that demographic and medical information relating to exclusion and inclusion criteria is readily available in most medical registries, this can allow for a more rapid assessment and recruitment of study participants (8). A recent systematic review found that 8 of 24 investigated registry-based randomised controlled trials exceeded their sample size requirements by using medical registries as a platform for participant recruitment (20). This platform for more rapid and complete study enrolment can help reduce current issues in which nearly half of studies require extensions related to participant recruitment, and many fail to achieve recruitment targets, impacting study power and preventing the ability to adequately answer the primary research question (11, 21, 22).

Another potential advantage of registry-based designs is the completeness of follow-up. Since registries continually collect medical information on patients, quick and complete long-term follow-up is possible without the need for tedious clinical reporting forms (8, 17, 23). Loss to follow-up is a significant barrier in standard clinical trials, and it is estimated that 60-89% of clinical trials are affected by missing follow-up and outcomes data, which can severely impact study power and quality (24, 25). The comprehensive and longitudinal nature of medical registries helps to prevent issues of missing outcomes data, as demonstrated in the TASTE trial, in which no patients were lost to follow-up (15).

The success of the TASTE trial and subsequent studies in the cardiology setting using the Swedish web-system for the enhancement and development of evidence-based care in heart disease according to recommended therapies (SWEDEHEART) registry as a platform for most key trial processes (e.g., recruitment, randomisation, data collection) is a testament to the

benefits of registry-based research. These studies have similarly demonstrated ease of recruitment, near-complete follow-up, limited missing data, and low cost (26-28). A recent systematic review further highlighted the advantages of registry-based trials, including shorter trial times, smaller carbon footprint, and lower participant burden (20). Registry-based research offers several unique and tangible benefits to the research process.

Despite registry-based research offering several benefits regarding cost, study quality, and feasibility, and despite their apparent success in fields such as cardiology, these study designs remain less common in oncology. A systematic review investigating registry-based randomised controlled trials published up until June 2020 found that only seven trials were conducted in the field of oncology (20). These studies used cancer registry and other registry data for participant enrolment, randomisation, and endpoint ascertainment, investigating the efficacy of interventions such as screening procedures on cancer incidence and smoking cessation on surgical outcomes (29-35). However, this systematic review used a narrow definition of registry-based trials, whereby they only included studies that used registry data for patient recruitment and ascertainment of at least one outcome measure. Thus, many more registry-based trials that use registries for only one aspect of their study design (e.g., outcome ascertainment only) likely exist. A study conducted in the UK investigating cancer mortality after long-term follow-up in the UK Collaborative Trial of Ovarian Cancer Screening (UKCTOCS) linked participants to national cancer and death registration datasets to evaluate outcomes (36). The study highlighted the linkage to national registries as a strength due to the completeness of follow-up afforded by this data acquisition method (36). Still, despite the success of these studies and other registry-based trials, this study design is not extremely common. In fact, among randomised controlled trials conducted in the UK between 2013 and

2018, only about 3% utilised routinely collected health data, including registry data, in their study designs (37). These findings are likely because, despite the several advantages of registry-based research, disadvantages remain.

Although registry-based research has several advantages over standard trials, there are also potential drawbacks. Missing data, limited endpoint selection, poor timeliness, and ethical and methodological challenges are just some of the disadvantages of registry-based trials (8, 9, 20). Some ethical challenges that may arise in the use of medical registries include issues of informed consent for patient screening and study inclusion, patient privacy, and data protection (8). Registry-based trials can also pose methodological challenges, including barriers to study blinding and limitations to study design, as trials are confined to data already routinely collected in the registry (8). However, probably the biggest concern is that the research is dependent upon the quality of registry data, which can vary significantly. Ensuring that registry data is complete, reliable, and timely is essential for the effective use of registry data in research.

1.2. Elements of Cancer Registry Data Quality

The quality of data available in medical registries can vary significantly based on the registry, the specific disease, and patient characteristics. The usefulness of a medical registry for research largely depends on data quality. For cancer registries specifically, four main elements of data quality have been identified: completeness, validity, comparability, and timeliness (38, 39). Completeness refers to how many of the diagnosed cancers within a given area are present in the corresponding cancer registry (39). In this sense, a high level of completeness is necessary to have an accurate picture of the cancer landscape in a given population. Occasionally, there is data that we are aware we do not know. For example, cancers of unknown primary do not have

their own staging system, so stage data for these cancers often indicate “not applicable.” Listings of “not applicable” would suggest that the data is missing for some known reason but are not incomplete. However, sometimes data fields are not completed for unknown reasons. These data are considered incomplete because we do not have a known cause to explain why this data point is not present. If there is a significant amount of incomplete data, this can pose problems to research and can limit the ability to investigate the research question.

Comparability relates to the ease with which cancer registry data can be compared and translated between groups and within groups over time (38). To achieve comparability, there should be clear definitions of all variables included in the registry, including incidence, prevalence, primary cancers, etc. Furthermore, standardised classification codes, such as the World Health Organisation’s (WHO) International Classification of Diseases (ICD) codes, should be used, and registries should utilise standardised methods of cancer staging. A high level of comparability allows for assessment and comparison both between different registries and within registries over time.

Validity refers to the accuracy of the data listed within a registry (38). In testing literature, it is important to distinguish between accuracy and precision. Accuracy refers to how well information reflects the accepted truth, whereas precision is concerned with the amount of variance between measurements of a variable. Regarding data validity for cancer registries, the focus is more concerned with accuracy and how well the data in the registry reflects the truth, where the true or accepted value is often what is found in the medical record or some other standardised dataset. Although there are no universal definitions of what constitutes data that is valid enough to be used in research, several studies have suggested appropriate cut-offs to consider. One author suggested that a discordance rate below 5% for each variable in a registry is

considered valid (40). This idea was further clarified by the same author in another publication, explaining that a completeness exceeding 95% with the number of missing registrations not exceeding 10% should also accompany the discordance rate below 5% for the data to be considered valid (41). Timeliness refers to the speed at which cancer data is accurately obtained and reported within the registry (38). There are no clear definitions for what constitutes timeliness of cancer registration. Still, it is often calculated as the lag time between the initial cancer diagnosis and when the cancer is reported in the registry. Accurate and timely data in a cancer registry is essential for its use in cancer research.

As one can imagine, these four factors, completeness, comparability, validity, and timeliness, exhibit trade-offs. An increase in the completeness and validity of a cancer registry often comes at the sacrifice of timeliness, as more thorough systems are put in place to ensure complete and valid data. The opposite is also true, in that more rapid data acquisition in the registry often comes at the cost of less complete and accurate data. Given that these factors are all critical for utilising cancer registry data in research, these qualities must be evaluated to assess the quality of the registry data.

1.3. Evaluating Cancer Registries in Other Countries

Given the differences between national cancer registries, each cancer database must be evaluated independently to explore its data quality. Many studies have been undertaken to assess the quality of cancer registry data in various registries around the world. In order to evaluate the data quality of central cancer databases, data in national registries is often compared to data obtained from a recent study or extracted from medical records.

1.3.1. Non-European Countries

1.3.1.1. Canada

A study in Canada evaluated the accuracy of kidney cancer data in one of the provincial/territorial cancer registries in the country. The study compared registry data regarding the pathology and diagnosis of renal cell carcinoma to pathology reports and included patients diagnosed between 2006 and 2010 (42). Among the 723 patients included, the study found a discordance rate of 15.4% between registry data and that found in the pathology report, and the most common reasons for discordance were a lack of information from the pathologist (45.5%) and a true coding error (32.3%) (42).

Another study looking at germ cell tumour histology and stage within the same provincial/territorial cancer registry was conducted on all patients diagnosed with a germ cell tumour between 2006 and 2015 (43). Data from the cancer registry was compared to patient medical charts. The study found a discrepancy rate of 5.2% in the location of the primary germ cell tumour, 7.0% in pathology, and 17.3% in stage. In the 17.3% of discordant stage cases, the registry had a lower stage than reported in the patient chart in 4.3% of the cases, while the remaining 11.4% had a higher stage indicated in the registry (43).

1.3.1.2. New Zealand

In an analysis of the New Zealand Cancer Registry (NZCR), staging data on primary breast cancers registered in the Waikato Breast Cancer Register (WBCR) were compared to those of the NZCR for cancers diagnosed between 1999 and 2011 (44). The study compared 2,662 patients diagnosed with primary invasive breast cancers and found a rate of missing stage data of 12.3% in the NZCR, but an overall concordance of 94.4% between the NZCR and WBCR (44). The study found that metastatic and locally invasive cancers tended to have higher

proportions of missing or inaccurate stage data (44). Several other factors were identified to have an influence on the likelihood of missing stage data, including older age, higher comorbidity score, mortality, and not undergoing therapeutic surgery (44). Overall, this led to an underestimation of the incidence of metastatic breast cancer in the NZCR by 21% (44).

Similar studies have looked at lung and colorectal cancer data in the NZCR. One study used data from an audit of secondary cancer management of lung cancer patients in New Zealand in 2004 to compare the NZCR data against audit cases from regional databases (45). The study found that 12% of the 565 audit cases were not registered in the NZCR (45). There are instances where cancer registries may aim not to register a cancer, such as in cases of a cancer recurrence, a second primary, or when a patient was diagnosed with cancer in a given region, but the person does not reside in that area. However, these are relatively infrequent occurrences and would almost certainly not account for the number of missed cases in each registry. In addition to missing cases, disease extent was available for only 58% of cases in NZCR, with similar findings to previous studies that disease extent data was more likely to be missing for those of older age and higher comorbidity or those with locally advanced disease (45). Similarly, medical record data from 642 colorectal cancer cases diagnosed between 1996 and 2003 were used to evaluate the quality of cancer data in the NZCR (46). This study found 95% accuracy for tumour site, 86% accuracy within tumour site along the colon, 83% accuracy for tumour grade, and 80% accuracy for tumour stage (46). These studies further highlight discrepancies in stage reporting in the NZCR.

Another study investigated the completeness and accuracy of childhood cancer registrations in the NZCR and the New Zealand Children's Cancer Registry (NZCCR). The study found overall high case completeness for the NZCR at 99.3% and for the NZCCR at 94.4%

(47). Meanwhile, the accuracy of the six data fields assessed was also high at 98.6% among the two datasets (47). This study illustrated an overall high level of completeness and accuracy of childhood cancers in the various New Zealand cancer registries.

1.3.1.3. United States of America

The Centres for Disease Control and Prevention (CDC) in the United States established the National Program of Cancer Registries (NPOCR), which oversees central cancer registries throughout the country. An analysis of the case completeness and data quality of NPOCR-supported cancer registries compared to medical records found an average case completion of 96.4% and data accuracy of 95% for lung and bronchus, colorectal, prostate, and female breast cancers diagnosed between 1998 and 2001 (48). The study found the lowest accuracy rate for stage, which had an average accuracy of 88.8% (48).

Another study evaluated the accuracy and completeness of melanoma of the skin, bladder, pancreatic, kidney and renal pelvis, and ovarian cancers diagnosed in 2018 at 17 central cancer registries funded by the NPOCR (49). The study found similarly high accuracy of cancer data compared to previous studies with an average of 98.0%, with the highest accuracy for kidney and renal pelvis data at 98.8% and the lowest for bladder cancer at 97.4% (49). Of the errors that were identified, 53% were in treatment information, followed by 27% in cancer identification, and 20% in stage and prognostic factors (49).

The National Cancer Database (NCDB) of the United States is another cancer registry that includes information on over 70% of cancers diagnosed throughout the country (50). One study compared the results of comparative effectiveness research in 141 randomized clinical trials to observational data obtained from the NCDB (51). This study found that 56% to 70% of

studies had concordant hazard ratios, while only 41% to 46% of studies had concordant P values, indicating that comparative effectiveness research using NCDB data often obtains results that are discordant with randomised controlled trials (51). Thus, endpoint selection and study type are important considerations before engaging in registry-based research using the NCDB.

1.3.2. European Countries

1.3.2.1. Denmark

A review of the evidence looking at the Danish Cancer Registry (DCR) reported an overall high level of missing tumour, nodes, and metastasis (TNM) stage information in the cancer registry (52). Notably, TNM data was dependent on the cancer site, with two-thirds of prostate cancer and over half of bladder cancer patients having incomplete staging information available in the registry (52). Demographic information suggested that older patients and those with high comorbidity were less likely to have complete staging information available in the DCR (52).

An analysis of breast cancer data in the DCR compared to clinical records demonstrated a high case ascertainment, in which no tumours were missing from the cancer registry (53). Additionally, concordance between the two data sources regarding diagnosis was high at 99% (53). However, stage concordance was lower, with crude staging discordance at 13% (53). Similar to other studies, stage discordance was higher for metastatic cancers and those with bilateral breast tumours (53). There was also a range in the concordance for treatment, with surgery showing a concordance of 95.7%, compared to only 72.7% for adjuvant chemotherapy (53).

Another study comparing prostate cancer data in the DCR with hospital records found overall high case registration and minimal differences between the two methods in the date of

diagnosis reported (54). However, there were inconsistencies in cancer stage; for example, only 57% of the cases of distant metastases found in the hospital records were recorded in the registry (54). There were also inconsistencies in treatments, with 33 patients reported as having radical prostatectomy in the DCR, yet only two cases were confirmed by the hospital records (54). Overall, studies of the DCR demonstrate high case ascertainment but variable concordance with medical records for different cancer variables.

1.3.2.2. Finland

The Finnish Cancer Registry (FCR) was evaluated by looking at the accuracy of colorectal cancer data in the FCR from patients diagnosed during a colorectal screening program between 2004 and 2012 compared to data extracted from the same patient records by two gastrointestinal surgeons (55). The study looked at the accuracy of tumour characteristics and treatment information. The study included 1,475 patients and found moderate agreement between the two data sources for clinical stage (k=0.74) and histopathology (k=0.72), high agreement for tumour location (k=0.87), and fair agreement for primary tumour removal (k=0.46) and chemotherapy (k=0.47). However, the study found an overall lower sensitivity for stage in the FCR in cancers that were metastatic and higher sensitivity for those that were localised or locally-advanced (55). There was more missing data in the FCR than in patient medical records for stage, surgical information, and adjuvant therapy status. Furthermore, 12% of patients lacked cancer report forms from clinicians, which are used to compile FCR data (55). Overall, the accuracy of the data quality of the FCR differed significantly depending on the specific cancer data explored, with the most obvious discrepancies between the two data sources in treatment information.

Another study looking at the timeliness of cancer registrations in the FCR compared registry data to 3,600 cancers diagnosed during the Alpha-Tocopherol Beta-Carotene Cancer Prevention (ATBC) Study between 1985 and 1997 (56). The study found that for most cancer sites, 95% of cancers were registered in the FCR within 0.9 months (56). However, there was variation in this timeliness amongst cancer types, with lung and pancreatic cancers experiencing longer delays to registration, with lung cancers taking an average of 1.7 years for 95% of cancers to be registered and 3.2 years for pancreatic cancers (56). Cancer site was concordant between the FCR and ATBC for 96% of cases found in the FCR, and only 0.8% of cancers found in the ATBC were not recorded in the FCR (56). This study demonstrated a high overall case ascertainment of the FCR, and differential timeliness of cancer registrations based on cancer site.

1.3.2.3. Norway

The Cancer Registry of Norway (CRN) was evaluated for the quality of data, including completeness, concordance, timeliness, and comparability, for the registration period of 1953-2005, but with particular focus on the 2001-2005 period (57). The study found appropriate comparability of the CRN, following international standards. The completeness of cancer registration data was found to be 98.8% (57). However, a relative lack of reporting was identified for haematological and central nervous system (CNS) malignancies (57). As for validity, 93.8% of the reported cancers between 2001 and 2005 were morphologically verified (57). There were cancer-specific differences in staging completion, whereby prostate cancer had the highest rate of missing stage data at 42%, compared to only 0.5% in female breast cancer cases (57). There was also a change in the timeliness of registration over the course of the study, with a median time between the date of diagnosis and registration of 525 days in 2001 and 261 days in 2005

(57). Overall, the authors concluded that the data in the CRN is reasonably comparable, accurate, complete, and timely.

1.3.2.4. Spain

One study looked at the completeness of clinical and TNM stage in all of the cancers in the cancer registry of Mallorca, Spain between 2006 and 2008 (58). The overall completion of T, N, and M stages was low at 48.6%, 36.5%, and 40.0%, respectively, while overall stage showed 37.9% completion (58). Stage completion exceeded 50% for certain cancer types, including lung, colon, ovary, and oesophagus (58). The investigation of demographic factors found no difference in completion of stage and TNM by gender but was lower for patients under 40 or over 80 years of age (58).

Evaluation of cancer data in the Spanish EURECCA Esophagogastric Cancer Registry (SEEGCR) from patients registered between 2014 and 2017 was compared to hospital records. Among 10,905 data items available for assessment, 4.7% were incorrect compared to medical records, and 0.3% were missing, indicating an overall high concordance and completion rate (59). Charlson comorbidity index showed the highest discordance rate at 12.4%, while the M stage was the most frequently missed at 2.4% (59).

1.3.2.5. Sweden

One study evaluated the completeness, timeliness, comparability, and validity of cancer data in the Swedish Colorectal Cancer Registry (SCRCR). The study found that between 2008 and 2015, there was an average completion of 98.5% and 98.8% for colon and rectal cancers, respectively, compared to the number of cancers registered in the Swedish Cancer Registry (60).

Regarding timeliness, 98% of cancers were registered within the SCRCR within 12 months of diagnosis in the final year of the study, and there were no differences in timeliness between colon and rectal cancers (60). The study used hospital records of 500 randomly selected cases from 2008 to compare validity and found that the average agreement between the two data sources was 90% (60).

Another study investigated the validity of surgical complications in a national cancer registry run by the National Board of Health and Welfare (NBH) in Sweden, a cancer registry run by the Regional Oncological Centres (ROC), and a local quality assurance system and compared the findings of each dataset to patient records. The study found the highest validity in surgical complications reported in the local quality assurance system (40). There was a high proportion of missed surgical complications in the NBH registry, at 69% and 64% for rectal and colonic cancers, respectively, and for the ROC registry, at 40% and 22%, respectively (40). The rate of missed surgical complications for cancers in the quality assurance system was only 7% (40). The majority of complications that were missed were wound infections rather than serious complications (40).

A study looking at the validity of cancer data in the Swedish Rectal Cancer Registry (SRCR) included 906 registered patients who were treated with major abdominal surgery between 1995 and 1997 (61). SRCR data was compared to medical records for 14 variables to assess validity. The study found a relatively high completion for most variables, including those regarding tumour factors, neoadjuvant therapy, surgical procedure, local radicality, and TNM stage (61). However, there was lower completion for anastomotic leakage, local recurrence, and distant recurrence, with missing data ranging from 13-38% for these variables (61). All variables showed substantial agreement between the registry and medical record data with correlation

coefficients between 0.82 and 1.00, except for rectal perforation, which showed moderate agreement with a correlation coefficient of 0.78 (61).

Another study evaluated the data quality of the Swedish National Register for Oesophageal and Gastric Cancer (NREV) for comparability, completeness, accuracy, and timeliness. The study found that the coding used in the NREV was consistent with international guidelines, indicating high comparability (62). Of the patients that were registered in the Swedish Cancer Registry from 2009 to 2013, 95.5% were also registered in NREV, indicating high completeness (62). There were 60 variables that were evaluated for accuracy between the registry and medical records among 400 randomly selected patients, resulting in a concordance rate of 91.1% between the two data sources (62). Exact agreement was lowest for smoking habits (72.6%), preoperative weight (74.8%), and clinical N category (79.8%) (62). Furthermore, the median time to registration was 3.9 months from diagnosis (62).

The data quality of the National Prostate Cancer Register (NPCR) for Sweden was also assessed. The registry demonstrated high comparability through its use of international reporting guidelines. The study found that among cancers registered in the Swedish Cancer Registry between 1998 and 2012, 98% were also registered within the NPCR (63). Furthermore, average completeness was estimated at 90% over 48 variables that were evaluated. In the final year of the study in 2012, 95% of cases were found to be registered within 12 months of the diagnosis, compared to only 80% in 2009 (63). Data from the registry was compared to hospital records, and it was found that there was an overall high level of concordance, with exact agreement for Gleason score, clinical local T stage, and androgen deprivation therapy at 97%, 83%, and 95%, respectively (63). The exact agreement was lower for certain variables, such as the main reason

for the initiation of work-up leading to prostate cancer diagnosis and the date of treatment decision, which were only 67% and 69%, respectively (63).

A study looking at the completeness of the Swedish Cancer Registry in 1998 found that according to hospital records, 3,429 of 42,010 individuals with cancer were not registered in the Swedish Cancer Registry (64). A subset of these 3,429 cases was reviewed to determine whether they should have been included in the registry, and the results found that about half of the cases were, in fact, missed (64). The study indicated a missed registration rate in the registry of about 3.7%, which indicates an overall high level of completeness (64). The study also found that soft tissue, nervous system, and blood cancers, including leukaemia and lymphoma, experienced higher rates of underreporting in the registry (64). Additionally, demographic factors played a role, as underreporting of cancer was more prevalent in patients over the age of 70 (64).

A validation of the National Breast Cancer Register (NBCR) for Sweden found that 99.9% of cancers diagnosed between 2010 and 2014 that were registered in the Swedish Cancer Registry were also registered in the NBCR (65). The registry had high comparability due to the use of international guidelines. The registry also showed good timeliness, with 98.5% of cancers being registered within 12 months of diagnosis (65). In assessing validity, 800 patients were randomly selected to have data re-abstracted from their medical records. The exact agreement between medical records and the NBCR was >90% for many of the variables examined but was notably less for some variables, including T classification (70.1%), the reason for no primary surgery (53.1%), the reason for reoperative axillary surgery (57.6%), and some histopathology and postoperative variables (65).

Another study looked at the completeness and accuracy of renal pelvic and ureteral cancers in the Swedish Cancer Registry, diagnosed between 1971 and 1998. There were 939

patients included, and 7.2% were found to be a false-positive inclusion in the registry, which was defined as an incorrect listing of the renal pelvis/ureter as the primary site, or a case that was not malignant (66). Additionally, there were 28 patients that were identified from previous studies that were missing in the registry (66).

1.3.3. United Kingdom

1.3.3.1. Northern Ireland

A comparison of Northern Ireland Cancer Registry (NICRD) data to that of general practitioner (GP) records found overall high agreement between the two data sources for 17,102 patients diagnosed with cancer between 1993 and 2010 (67). The study found that the completeness of the registry was 99.9%, with only 0.08% of cancers identified in GP records not found in the registry, and only 0.02% had a diagnosis date in the registry that varied by more than two weeks of what was found in GP records (67). Furthermore, only two discrepancies in tumour type and three discrepancies in tumour status were found between the two data sources (67). The study indicated an overall high concordance and completeness of the NICRD.

1.3.3.2. Scotland

Evaluation of Scottish cancer registration data was conducted in the 1990s, when data from 2,200 registered cancers were compared to data from medical records (68). The study found that, in addition to discrepancies in identifying information in 3.5% of cases, there was also discordance between registry and medical record data for ICD-9 codes and ICD-O morphology codes at rates of 5.4% and 14.5%, respectively (68). It was deemed that serious discrepancies (e.g., wrong year of diagnosis, little to no evidence of neoplasm, not a Scottish resident at the time of diagnosis, reclassified from malignant to benign, etc.) occurred in only 2.8% of cases

(68). As such, the authors concluded that the Scottish cancer registration data demonstrated a high level of accuracy. A similar study was carried out by the same author, which compared cancer registration data to medical record data for 309 lung cancer registrations from the year 1990 (69). Again, the registry demonstrated a high level of accuracy, with discrepancies in identifying information in 5.2% of patients and with discrepancies in ICD-9 and ICD-O morphology codes at 4.2% and 15.5%, respectively (69).

1.4. Central Cancer Data in England and Wales

1.4.1. England

In England, the National Cancer Registration and Analysis Service (NCRAS) is responsible for the National Cancer Registration Dataset (NCRD), which is the country's national cancer registry. The registry has population-based national information on cancer diagnoses across England since 1971 (70). NCRAS was part of Public Health England, which was formed in 2013 before NCRAS was transferred to NHS Digital in 2021 (70). NHS Digital ultimately merged with NHS England in 2023, which is now the body that oversees NCRAS. Before this, the national cancer dataset was the National Cancer Data Repository, and regional registry data was used. NCRAS collects data from the National Health Service (NHS) from multiple sources to provide information on malignant and pre-malignant neoplasms in England (70). All collected cancer data is then reviewed by cancer registration officers (CROs) who evaluate the findings and seek out additional information if needed, either by reviewing reports and electronic health records or by corresponding with primary or secondary care (70). New registrations are evaluated by two CROs in order to ensure the quality and accuracy of the data (70). Demographic information, tumour data, treatments, and outcomes are all collected and

reported in NCRD. As the national cancer registry for England, NCRD collects data on all cancers diagnosed in the country. Although patients may choose to opt out of cancer registration, this occurs in less than one in 10,000 patients, making this dataset highly representative of the cancers diagnosed in the country (70). Given the rigorous registration process, large coverage, and comprehensive data collection, the NCRD is the gold standard cancer registry for England.

In addition to the NCRD, England has a rapid cancer registration dataset (RCRD) that supports near real-time analysis of cancer data in England and is quicker than the gold standard registration process (71, 72). This dataset includes cancer data from 2018 onwards and was developed to support ongoing public health responses during the COVID-19 pandemic (71, 72). The dataset is formed by the rapid processing of the best available data at the time, largely from the Cancer Outcomes and Services Dataset (COSD) and less so from the Cancer Waiting Times (CWT) and Hospital Episodes Statistics (HES) datasets, but does not use the multitude of data sources or registration checks utilised in the NCRD registration process (72, 73).

As the registration process and quality assurance checks differ significantly between the RCRD and NCRD datasets, data from these two sources will not necessarily match. A comparison of RCRD and NCRD data from April to September 2018 found that 11.5% of cancers identified in NCRD were missed in the RCRD, while 6.1% were false registrations that were identified in RCRD but not linked to a cancer in NCRD (73). Further analysis found that error rates varied based on cancer sites, with higher errors for bone, soft tissue, and unknown tumours, but lower error rates for sites including breast and prostate (73). Monthly reports are released that highlight the data quality and comparisons between the RCRD and NCRD. Still, the RCRD provides more up-to-date cancer registration data compared to NCRD, whereby cancer

data in RCRD is often available 3-4 months following diagnosis, while finalised NCRD data can take 21-24 months following diagnosis (74).

1.4.2. Wales

The Welsh Cancer Intelligence and Surveillance Unit (WCISU) is governed by Public Health Wales and serves as the NCRD-equivalent and national cancer registry for Wales. WCISU contains cancer registration data on cancers dating back to 1972, with over 686,000 records included (75, 76). Data on cancer treatments began in 1995, and staging data for malignant melanoma, breast, colorectal, and cervical cancers started in 2001, before staging for all cancers was made available in 2010 (75, 76). Cancer information is collected from several sources, including NHS organisations, surveys, research, and more. The registry uses internal processes and checks like NCRD to ensure the accuracy of cancer registration. Like NCRD, WCISU uses an opt-out registration system.

Digital Health and Care Wales (DHCW) is part of NHS Wales, and aims to use data and digital solutions to improve clinical care and empower patients (77). DHCW was developed in 2021 and replaced the NHS Wales Informatics Service and includes health records for the entire population of Wales (77). The information provided in DHCW records are utilised by NHS staff in patient care (77). In addition to supporting patient care, DHCW has the capacity for bespoke data requests, which has the potential to support research activities. Although not a cancer registry, DHCW, like RCRD, can provide administrative real-time data about cancer diagnoses.

1.5. Evaluation of English and Welsh cancer registries

Despite significant research into other national registries worldwide, the cancer registries of England and Wales have been under-scrutinised in comparison. One study investigated the data quality of English bowel cancer registrations between 1996 and 2004 and found that completion of the Dukes stage was 60% for all of the registrations evaluated (78). However, this and other studies were conducted before the launch of NCRAS in 2013. Therefore, at the time of this study, cancer registrations were split up regionally rather than under one national service, limiting the value of this comparison for the current state of the English cancer registry (78, 79). A more recent study looked specifically at the completeness and agreement between NCRAS and data abstracted from medical records for the Cluster randomised trial of prostate specific antigen testing for Prostate cancer (CAP) trial for prostate cancer TNM stage and Gleason grade (80). The study found that completeness for both TNM stage (29.9% vs. 67.6%) and Gleason grade (41.2% vs. 76.7%) was lower in NCRAS than in CAP and that agreement between the two datasets was high for Gleason grade but lower for TNM stage (80).

There have been no studies to date that have evaluated cancer data from the national cancer registry of Wales. However, the United Kingdom and Ireland Association of Cancer Registries (UKIACR) publishes reports on performance indicators of the cancer registries in the United Kingdom and Ireland. The 2021 report on 2019 data reported high tumour information completion for English and Welsh registries at 97.0% and 96.9%, respectively, and staging completion at 71.8% and 81.1%, respectively (81). Although these studies and data provide insights into cancer registry data quality in England and Wales, it remains true that no recent study has compared data from locally collected data in a cancer study to that of the national

registries of England and Wales across a wide range of cancers to compare the completeness, accuracy, and timeliness between these data sources.

1.6. Study Aims

I aimed to compare diagnostic cancer data obtained on-site during a prospective cohort study in England and Wales with data from central cancer databases in England and Wales to assess the feasibility of using these central datasets in research by comparing data completeness, concordance, and timeliness between the different data sources.

2. Methods

2.1. SYMPLIFY

The SYMPLIFY study was a prospective observational cohort study (ISRCTN10226380) that evaluated a multi-cancer early detection (MCED) test in symptomatic patients in England and Wales. The study recruited 6,238 participants across 44 hospitals in England and Wales between July 7 and November 30, 2021, who were referred by their general practitioner to a rapid diagnostic centre (RDC) or gynaecological (gynae), lung, upper gastrointestinal (GI), or lower GI two-week-wait (2WW) pathway for rapid investigation of symptoms suggestive of potential cancer (82). A total of 387 patients were excluded due to sample errors, participant withdrawal, inability to draw blood for the test, or recognition of ineligibility after enrolment, leaving 5,851 clinically evaluable patients (82). A further 390 were excluded: 376 because of no MCED test result and 14 due to missing information about the final diagnosis. Thus, 5,461 patients were included in the final cohort (82). Patients in the study also consented to the study team to retrieve centrally held cancer registry data. An exploratory objective of SYMPLIFY was to evaluate the quality of cancer data collected from central registries by assessing the proportion of completed data fields, the concordance between study data and central registry data, and the difference in timeliness between the SYMPLIFY and central datasets.

Hospital sites were asked to review the hospital records for all participants at 3 months following study enrolment. Those that were unresolved at the 3-month mark underwent an additional review at 9 months following enrolment to identify and record any cancers identified locally using the hospital record and input cancer information into a secure online case report form (CRF). Notably, if a cancer was diagnosed at the 3-month follow-up, no further review was undertaken. Sites were encouraged to use the best clinical evidence available at the time of

completing the CRF.

2.2. Datasets

On-site data cuts from the SYMPLIFY hospital sites were retrieved monthly from April 2022 to January 2023 and included data from both English and Welsh patients. To allow for comparisons between site and registry datasets, SYMPLIFY was split up into SYMPLIFY-England and SYMPLIFY-Wales for completeness and concordance purposes. Linked registry data were collected from WCISU and NCRD from Wales and England, respectively. Rapid cancer data were collected from DHCW in Wales and RCRD in England. Monthly data cuts were available from RCRD from April 2022 to September 2023, NCRD from November 2022 to September 2023, DHCW from September 2022 to May 2023, and a single data cut was provided from WCISU in July 2023. Notably, the datasets obtained from the English registries were specifically curated for the SYMPLIFY study due to resourcing and early engagement in the study design. The Welsh registries were engaged later in the study design. This later engagement, compounded with resource limitations, led to WCISU only being able to provide data on patients with confirmed cancer cases, according to DHCW. Data from other registries was obtained on all 5,461 patients enrolled in the SYMPLIFY study, regardless of cancer status reported in SYMPLIFY or in other datasets.

Given that patients in SYMPLIFY were followed-up only for a maximum of 9 months following their enrolment in the study, cancers diagnosed more than 9 months after enrolment were excluded from the analysis. As such, each data cut included the most up-to-date cancer data available for cancers diagnosed within 9 months of study enrolment.

2.3. Cancer Data

As SYMPLIFY was a diagnostic study evaluating the performance of an MCED test, information on each cancer was focused on diagnostic data. There were four diagnostic data fields evaluated for completeness, concordance, and timeliness: International Classification of Diseases, Tenth Revision (ICD-10) code, International Classification of Diseases for Oncology, 3rd Edition (ICD-O-3) morphology code, TNM classification, and overall stage at diagnosis.

2.3.1. ICD-10 Code

ICD-10 codes are international classification codes that classify medical conditions and symptoms (83). In terms of cancer diagnoses, ICD-10 codes indicate the specific site of a cancer. ICD-10 codes demonstrate exceptional anatomical specificity and, as such, they demonstrate significant overlap in more general locations. For example, C18.0-C18.8 all indicate a cancer of the colon but vary in where the cancer is located along the bowel (e.g., C18.2 corresponds to a malignant neoplasm of the ascending colon, while C18.3. corresponds to a malignant neoplasm of the hepatic flexure) (84). As such, broad groupings of ICD-10 codes by related cancer sites were used to compare concordance between datasets, according to those groups previously outlined in the SYMPLIFY study (82). If a code that did not correspond with an ICD-10 code was listed, the cancer site was listed as "Unknown Code." Only ICD-10 codes starting with "C", indicating malignant neoplasms, or D45-D46.9, D47.3, D47.4 were included from all datasets. ICD-10 codes of C44 were excluded, as non-melanoma skin cancers were not included in the original SYMPLIFY study (82).

2.3.2. ICD-O-3 Code

ICD-O-3 codes are more specific classification codes used to classify cancers based on site, histology, and tumour behaviour (85). ICD-O-3 codes indicate a topographical code, indicating where a given cancer has emerged inside the body, and a histological code, identifying the morphology of the malignancy. Given that the topographical component of an ICD-O-3 code uses similar coding to the ICD-10 code, only the histological component was used in comparisons among equivalent cancers between datasets. The histological code contains a four-digit code to indicate the histological term, with a fifth number separated by a backslash to indicate the behaviour code (i.e., benign, malignant, carcinoma in situ, or a neoplasm of uncertain malignancy) (86). Only cancers with a behaviour code of 3, corresponding to a malignant neoplasm, were included in the analysis. Much like ICD-10 codes, there is significant overlap in ICD-O-3 coding, and so broader morphology groupings were developed based on those outlined in the International Classification of Diseases for Oncology 3rd Edition, first revision (Appendix p. 1-22) (87). If an ICD-O-3 code that did not correspond with a code in the third edition was listed, the second edition was referenced (87). If a corresponding ICD-O-3 code could not be found in either edition, it was classified as "Unknown Code".

2.3.3. Stage

Stage was defined as the overall cancer stage I-IV at diagnosis that was assigned to a given cancer. Any stage 0 cancers, corresponding to carcinoma in situ, were excluded from all datasets, as they were in the original SYMPLIFY study (82). The different datasets varied in how stage was reported: some reported sub-staging information (e.g., stage IIa, IIb), while others used broader stage groupings only (i.e., stage I, II, III, IV). Thus, to allow for comparisons between

datasets, only the broad numerical stage component was considered for concordance between datasets. The SYMPLIFY dataset used one of the following staging systems to report the overall stage: Union for International Cancer Control (UICC) 8th edition, European Neuroendocrine Tumour Society (ENETS), International Federation of Gynaecology and Obstetrics (FIGO), Ann Arbor, Binet, or other. WCISU used UICC 8th edition, FIGO, and Ann Arbor staging systems, while NCRD used UICC 8th edition, Ann Arbor, ENETS, FIGO, International Staging System (ISS), and Binet staging systems. The staging systems for RCRD and DHCW were not indicated.

2.3.4. TNM Classification

TNM staging was evaluated as a complete variable comprised of a combined TNM stage variable, as well as individual stage variables. All datasets that had TNM data included numeric and letter components in the stages, and therefore, both the numeric and letter components were considered. The systems used to record TNM stage were the same as outlined in section 2.3.3. for SYMPLIFY. The T, N, and M staging variables used for NCRD and WCISU only utilised the UICC 8th edition system. While other variables exist in the NCRD and WCISU datasets to incorporate other staging systems (e.g., FIGO), these were not included in the study. TNM staging information was not included in DHCW and was not available as a distinct variable in RCRD. Thus, the RCRD and DHCW datasets were excluded from TNM analyses. TNM stages were compared regardless of the staging system used in each dataset.

2.4. Analysis

All analyses were completed in Stata version 17.0 and figures were prepared in GraphPad Prism Version 9.4.1.

2.4.1. Completeness

The completeness of each data field was determined based on the number of cancers present in the dataset at that data cut, which had the given data point completed. Completeness of cancer data was reported as the number and proportion of cancers with available data for each data field among all cancers in that dataset per data cut with exact binomial confidence intervals. For TNM classification, the data field was considered complete if there was a non "N/A" entry for all of T category, N category, and M category. For the remaining data points, a blank field was considered incomplete. A field listed as "Uncertain" was considered complete, as "Uncertain" was often inputted in cases where no staging system was available, such as for cancers of unknown primary. Additionally, in cases where a patient could not undergo further staging investigations, "Unknown" was often entered.

In addition to the completeness of cancer data fields, I also investigated the completeness of cancer registrations. The number of cancers reported in SYMPLIFY but not the respective registry, and vice versa, was tabulated. An analysis of the time between enrolment and diagnosis was carried out for all cancers that were reported in SYMPLIFY but not in the respective registry to determine the number and percentage of cancers that were diagnosed within the mandatory ≤3 months post-enrolment follow-up period, >3 months and ≤6 months post-enrolment, and >6 months and ≤9 months post-enrolment. Additionally, cancers that were reported in SYMPLIFY but not NCRD were investigated by liaising with NCRAS to determine why they were not

included in the registry. I was unable to obtain the same information about missing cancers in the WCISU or rapid cancer datasets.

The cancer sites of those that were reported in SYMPLIFY but not the respective registries were identified to determine trends in missed cancer registrations. These were presented as the number and percentage of the total missing cancers in the registries and as the number and percentage of the total cancers identified at that site in the SYMPLIFY dataset. Additionally, cancers that were reported in NCRD and WCISU but not in SYMPLIFY-England and SYMPLIFY-Wales, respectively, were also investigated for trends in missed cancers by cancer site. These were presented as the number and percentage of missing cancers in the SYMPLIFY dataset and as the number and percentage of the total cancers identified at that site in the registry dataset.

2.4.2. Concordance

The concordance of data fields was assessed for cases when a cancer was recorded for the same patient in the SYMPLIFY and respective registry datasets. Patients with multiple cancers were dealt with on a case-by-case basis. Concordance was defined as a cancer having the same data field value inputted in both datasets. Concordance between datasets was determined by merging the two datasets on both the participant identification number and the given data field. Concordance was calculated as a percentage of the number of cancers present in both datasets that had the given data field completed in both datasets to avoid reporting a misleading discordance rate inflated by missing values. The number and proportion of concordant cancers were reported with exact binomial confidence intervals.

When possible, data cuts from SYMPLIFY and central databases from equivalent time points were compared to calculate concordance. Data cuts from central databases beyond January 2023 were compared to the final January 2023 SYMPLIFY data cut. For ICD-10 codes, concordance was based on the broad cancer site groupings. For ICD-O-3, concordance was determined based on an exact match based on the 4-digit histology code and separately again based on the broader morphology groupings (Appendix p. 1-22). For cancer stage, concordance was calculated based on a match between the broad numerical stage (i.e., I-IV) with no sub-staging information. For TNM classification, concordance was calculated based on a perfect match between combined TNM stage variables. Further analyses were conducted to compare the concordances for individual T, N, and M categories.

2.4.3. Timeliness

Timeliness was determined by evaluating the evolution of completeness and concordance over time by comparing each monthly data cut to the final data cut available for each dataset. With only one WCISU data cut available, the timeliness of completeness and concordance could not be investigated. The evolution of completeness was determined as the number and percentage of cancers with the data field complete at the given time point relative to the total number of cancers with the data field complete at the final time point for that dataset, according to the same analysis that was used to investigate completeness, outlined in section 2.4.1. The timeliness of completeness was expressed as the number and percentage of data fields recorded at each monthly data cut compared to the final data cut. In addition to looking at the completeness of data fields, the completeness of cancer registrations was also investigated by

determining the number and percentage of cancers registered in each data cut relative to the final data cut.

The evolution of concordance was determined by merging each monthly data cut with the final data cut available for a given dataset and assessing the number of concordant cases between the monthly and final data cuts for each data field. The data was reported as the number and proportion of concordant cases each month, compared to the number of cancers with the data field completed in the final data cut.

2.4.4. Completeness by Cancer Site and Referral Pathway

The completeness by cancer site and referral pathway was analysed by first determining the number of cancers present for each cancer site and referral pathway in each dataset in the last available data cut. The completeness was then calculated as the number and percentage of cancers at a given site or pathway that had the given data field completed, according to the same analysis that was used to investigate completeness, outlined in section 2.4.1. Only cancer sites that had at least ten cancers present at the last data cut available were included in the analysis.

2.4.5. Analysis of Discordant Cases

To investigate discordant cases, the differences in cancer site (ICD-10) and broad morphology groupings (ICD-O-3 groupings) between the final data cuts for each dataset were summarised. For TNM stage, I compared the overall concordance between the combined TNM stage to the concordances between individual T, N, and M categories to better elucidate the discrepancies seen amongst this variable. For stage, I compared discordant cases at each time

point to determine whether the stage listing was reported as higher in the registry or in SYMPLIFY in each case.

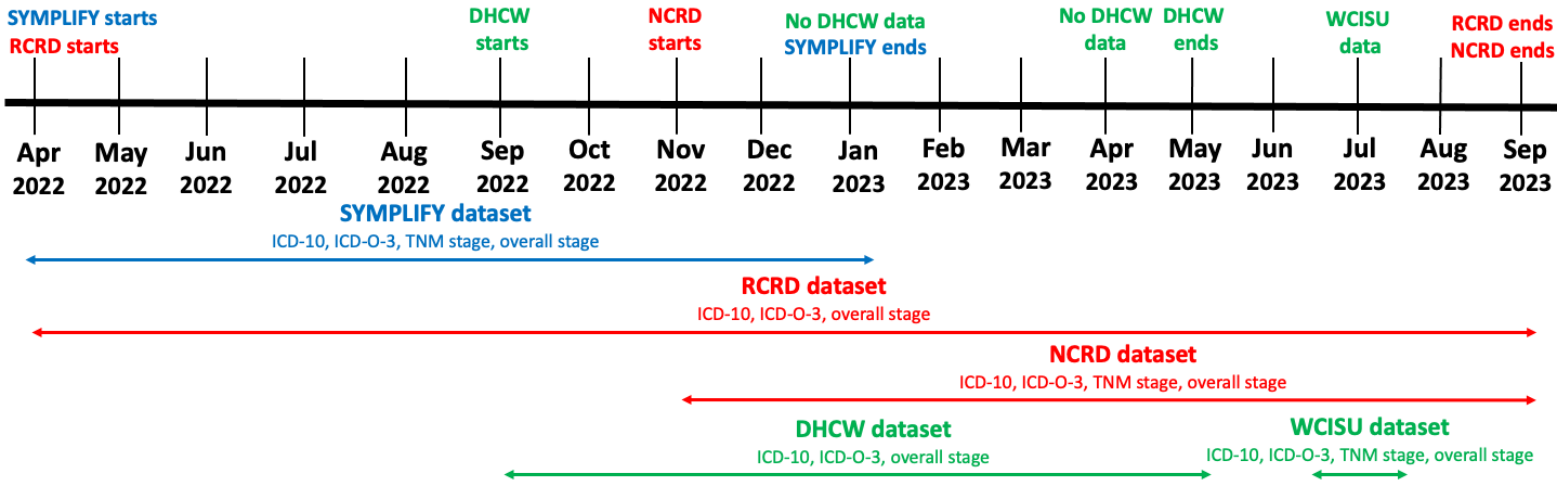
2.5. Outcomes

The primary outcome was a descriptive analysis comparing the completeness, concordance, and timeliness of completeness and concordance between cancer data collected on-site during the SYMPLIFY study and from the central cancer registries in England and Wales. Secondary outcomes included a summary of discordant cases between SYMPLIFY and the registry datasets and the completeness of cancer data by cancer type and referral pathway.

3. Results

Of the 5,461 participants included in the SYMPLIFY study, 4,370 (80.0%) were recruited in England, and 1,091 (20.0%) were recruited in Wales. Females accounted for 66.1% of all participants, and the median age at registration was 61.9 years (IQR 53.4-73.0) (82). Data cuts were available from April 2022 to January 2023 for SYMPLIFY, April 2022 to September 2023 for RCRD, November 2022 to September 2023 for NCRD, September 2022 to May 2023 for DCHW, and a single data cut was available for WCISU in July 2023. Figure 1 outlines the timings of the data cuts and data fields available from each dataset.

Figure 1. Timeline of the data cuts and data fields available for each dataset.



By the final data cut for each dataset, there were 259 cancers recorded within 9 months of enrolment among 250 participants in SYMPLIFY-England, 121 (118) in SYMPLIFY-Wales, 226 (221) in RCRD, 291 (276) in NCRD, 122 (118) in DHCW, and 112 (108) in WCISU (Figure 2, Table 1). Due to a data issue at one of the England hospital sites, there was a temporary drop in cancers and, consequently, a drop in completeness, concordance, and timeliness in the RCRD and NCRD datasets in the May 2023 data cut.

Figure 2. Number of cancers recorded in each dataset for the SYMPLIFY cohort over time.

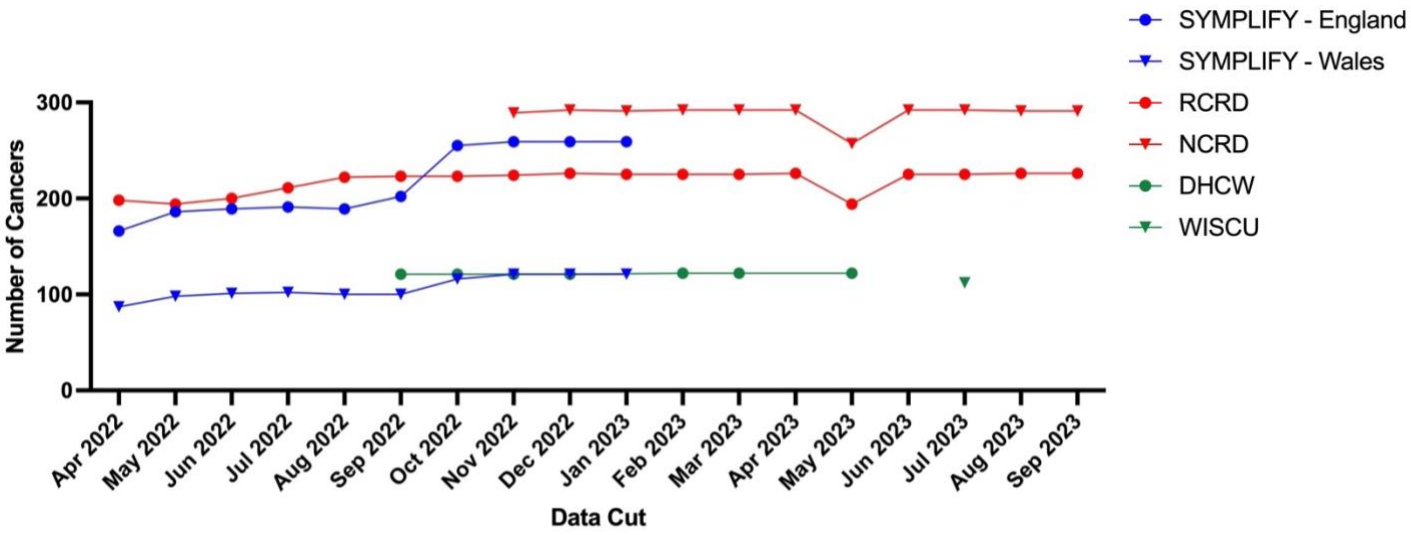


Table 1. Cancers recorded in each dataset for the SYMPLIFY cohort at each time point. Proportions are derived from the 4,370 participants recruited in England for RCRD, NCRD, and SYMPLIFY-England datasets, and the 1,091 participants recruited in Wales for DHCW, WCISU, and SYMPLIFY-Wales.

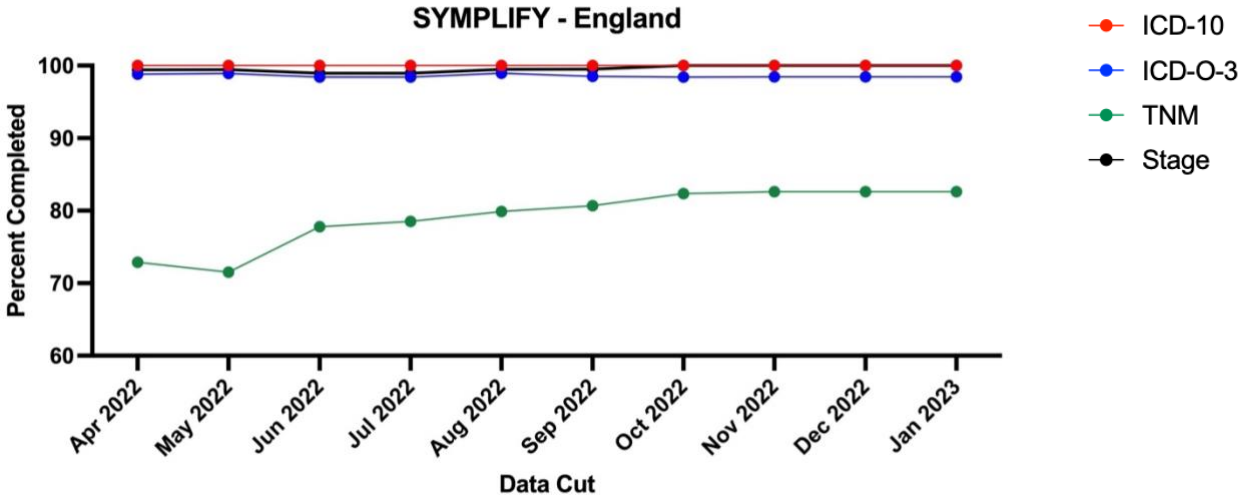
Data Cut	Dataset n cancers in dataset (n patients with cancer, % total number of patients)					
	SYMPLIFY		RCRD	NCRD	DHCW	WCISU
	England	Wales	England	England	Wales	Wales
Apr 2022	166 (163, 3.7)	87 (84, 7.7)	198 (195 4.5)	-	-	-
May 2022	186 (183, 4.2)	98 (94, 8.6)	194 (191, 4.4)	-	-	-
Jun 2022	189 (186, 4.2)	101 (96, 8.8)	200 (197, 4.5)	-	-	-
Jul 2022	191 (186, 4.2)	102 (97, 8.9)	211 (208, 4.8)	-	-	-
Aug 2022	189 (184, 4.2)	100 (95, 8.7)	222 (218, 5.0)	-	-	-
Sep 2022	202 (197, 4.5)	100 (95, 8.7)	223 (219, 5.0)	-	121 (117, 10.7)	-
Oct 2022	246 (244, 5.6)	116 (111, 10.2)	223 (219, 5.0)	-	121 (117, 10.7)	-
Nov 2022	259 (250, 5.7)	121 (118, 10.8)	224 (220, 5.0)	289 (275, 6.3)	121 (117, 10.7)	-
Dec 2022	259 (250, 5.7)	121 (118, 10.8)	226 (222, 5.1)	292 (278, 6.4)	121 (117, 10.7)	-
Jan 2023	259 (250, 5.7)	121 (118, 10.8)	225 (221, 5.1)	291 (277, 6.3)	-	-
Feb 2023	-	-	225 (221, 5.1)	292 (277, 6.3)	122 (118, 10.8)	-
Mar 2023	-	-	225 (221, 5.1)	292 (277, 6.3)	122 (118, 10.8)	-
Apr 2023	-	-	226 (221, 5.1)	292 (277, 6.3)	-	-
May 2023	-	-	194 (190, 4.3)	257 (243, 5.6)	122 (118, 10.8)	-
Jun 2023	-	-	225 (220, 5.0)	292 (277, 6.3)	-	-
Jul 2023	-	-	225 (220, 5.0)	282 (277, 6.3)	-	112 (108, 9.9)
Aug 2023	-	-	226 (221, 5.1)	291 (276, 6.3)	-	-
Sep 2023	-	-	226 (221, 5.1)	291 (276, 6.3)	-	-

3.1. Completeness

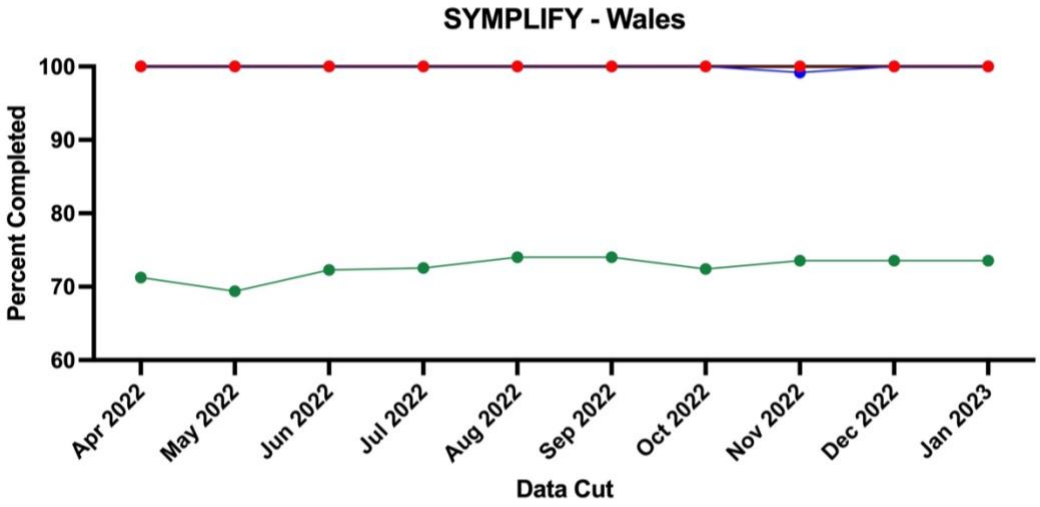
ICD-10 completeness was 100% at the final time point for each dataset (Figure 3, Table 2). ICD-O-3 morphology completeness varied amongst the different datasets, whereby for SYMPLIFY-England, SYMPLIFY-Wales, RCRD, NCRD, DHCW, and WCISU, the completeness was 98% (95% CI = 96%-100%, n=255), 100% (97%-100%, 121), 100% (98%-100%, 226), 100% (99%-100%, 291), 84% (76%-90%, 102), and 100% (97%-100%, 112), respectively (Table 2). Stage completeness was high for SYMPLIFY-England, SYMPLIFY-Wales, NCRD, and WCISU, at 100% (95% CI = 99%-100%, n=259), 100% (97%-100%, 121), 100% (98%-100%, 290), and 100% (97%-100%, 112), respectively, but lower for RCRD and DHCW at 73% (66%-78%, 164) and 43% (34%-53%, 53), respectively (Table 2). Among the four datasets that included TNM staging, the completeness was 83% (95% CI = 77%-87%, n=214), 74% (65%-81%, 89), 76% (71%-81%, 222), and 77% (68%-84%, 86) for SYMPLIFY-England, SYMPLIFY-Wales, NCRD, and WCISU, respectively (Table 2). There was a demonstrated increase in final staging completeness in the gold standard NCRD dataset (100%) compared to the rapid English RCRD dataset (73%). The Welsh registries demonstrated a similar pattern whereby final stage completeness in WCISU (100%) was much greater than in DHCW (43%). Still, the completeness of RCRD, NCRD, and DHCW datasets demonstrated consistency over time (Figure 3, Table 2).

Figure 3. Completeness (%) of data fields over time for panel (a) SYMPLIFY-England, (b) SYMPLIFY-Wales, (c) RCRD, (d) NCRD, and (e) DHCW and WCISU. ICD-10 and ICD-O-3 completeness all around 100% in SYMPLIFY, RCRD, NCRD, and WCISU, leading to overlapping lines.

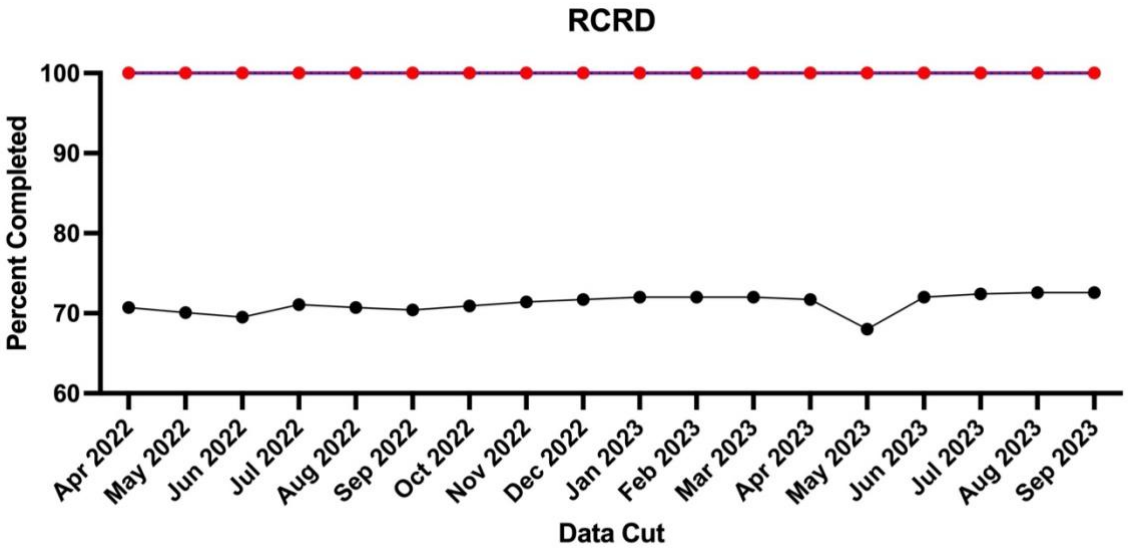
a)



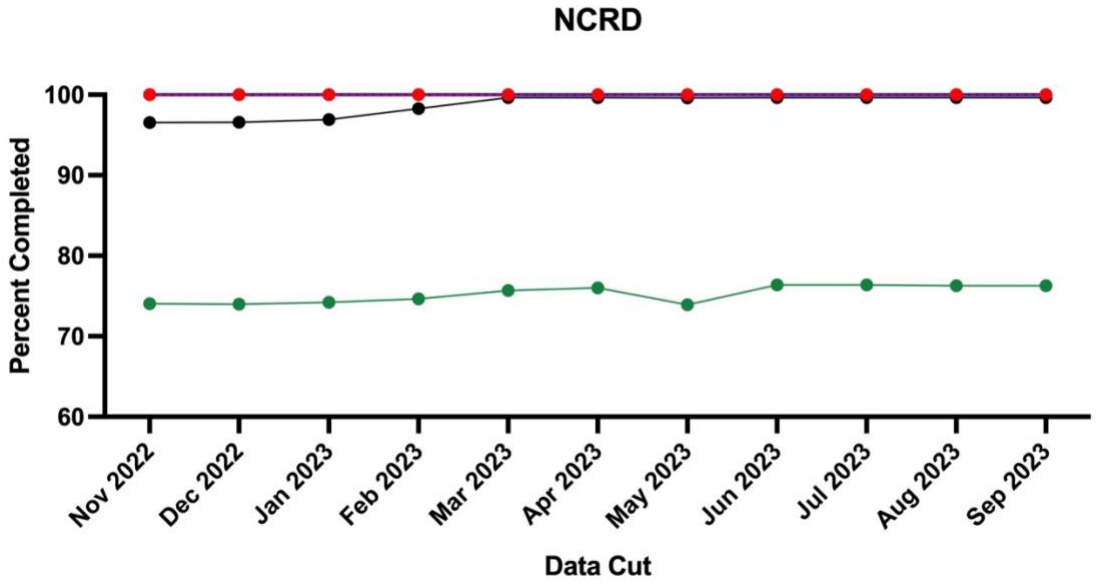
b)



c)


RCRD
Percent Completed
100
90
80
70
60
Apr 2022
May 2022
Jun 2022
Jul 2022
Aug 2022
Sep 2022
Oct 2022
Nov 2022
Dec 2022
Jan 2023
Feb 2023
Mar 2023
Apr 2023
May 2023
Jun 2023
Jul 2023
Aug 2023
Sep 2023
Data Cut

d)


NCRD
Percent Completed
100
90
80
70
60
Nov 2022
Dec 2022
Jan 2023
Feb 2023
Mar 2023
Apr 2023
May 2023
Jun 2023
Jul 2023
Aug 2023
Sep 2023
Data Cut

e)

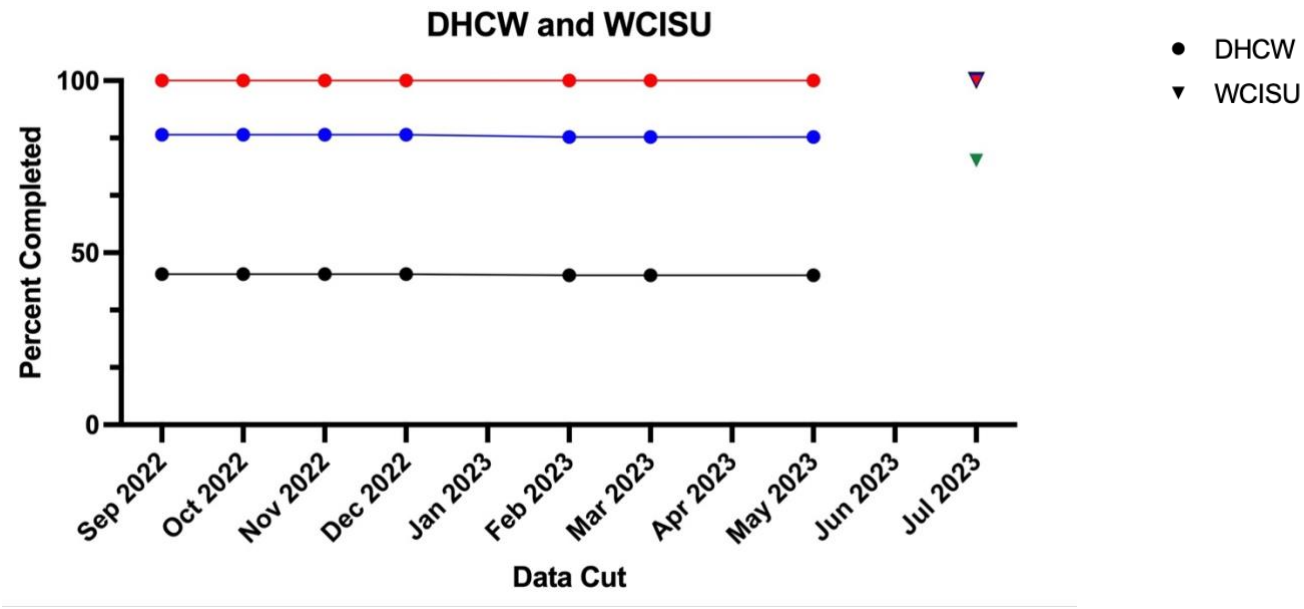

DHCW and WCISU
DHCW
WCISU
Percent Completed
100
50
0
Sep 2022
Oct 2022
Nov 2022
Dec 2022
Jan 2023
Feb 2023
Mar 2023
Apr 2023
May 2023
Jun 2023
Jul 2023
Data Cut

Table 2. Proportion (%) of cancers (n) recorded in each dataset at each time point with complete data for (a) ICD-10, (b) ICD-O-3, (c) TNM stage, and (d) stage.

a) ICD-10

Data Cut	SYMPLIFY		RCRD % (95% CI, n)	NCRD % (95% CI, n)	DHCW % (95% CI, n)	WCISU % (95% CI, n)
	England % (95% CI, n)	Wales % (95% CI, n)				
Apr 2022	100 (98-100, 166)	100 (96-100, 87)	100 (98-100, 198)	-	-	-
May 2022	100 (98-100, 186)	100 (96-100, 98)	100 (98-100, 194)	-	-	-
Jun 2022	100 (98-100, 189)	100 (96-100, 101)	100 (98-100, 200)	-	-	-
Jul 2022	100 (98-100, 191)	100 (96-100, 102)	100 (98-100, 211)	-	-	-
Aug 2022	100 (98-100, 189)	100 (96-100, 100)	100 (98-100, 222)	-	-	-
Sep 2022	100 (98-100, 202)	100 (96-100, 100)	100 (98-100, 223)	-	100 (97-100, 121)	-
Oct 2022	100 (99-100, 255)	100 (97-100, 116)	100 (98-100, 223)	-	100 (97-100, 121)	-
Nov 2022	100 (99-100, 259)	100 (97-100, 121)	100 (98-100, 224)	100 (99-100, 289)	100 (97-100, 121)	-
Dec 2022	100 (99-100, 259)	100 (97-100, 121)	100 (98-100, 226)	100 (99-100, 292)	100 (97-100, 121)	-
Jan 2023	100 (99-100, 259)	100 (97-100, 121)	100 (98-100, 225)	100 (99-100, 291)	-	-
Feb 2023	-	-	100 (98-100, 225)	100 (99-100, 292)	100 (97-100, 122)	-
Mar 2023	-	-	100 (98-100, 225)	100 (99-100, 292)	100 (97-100, 122)	-
Apr 2023	-	-	100 (98-100, 226)	100 (99-100, 292)	-	-
May 2023	-	-	100 (98-100, 194)	100 (99-100, 257)	100 (97-100, 122)	-
Jun 2023	-	-	100 (98-100, 225)	100 (99-100, 292)	-	-
Jul 2023	-	-	100 (98-100, 225)	100 (99-100, 292)	-	100 (97-100, 112)
Aug 2023	-	-	100 (98-100, 226)	100 (99-100, 291)	-	-
Sep 2023	-	-	100 (98-100, 226)	100 (99-100, 291)	-	-

b) ICD-O-3 morphology code

Data Cut	SYMPLIFY		RCRD % (95% CI, n)	NCRD % (95% CI, n)	DHCW % (95% CI, n)	WCISU % (95% CI, n)
	England % (95% CI, n)	Wales % (95% CI, n)				
Apr 2022	99 (96-100, 164)	100 (96-100, 87)	100 (98-100, 198)	-	-	-
May 2022	99 (96-100, 184)	100 (96-100, 98)	100 (98-100, 194)	-	-	-
Jun 2022	98 (95-100, 186)	100 (96-100, 101)	100 (98-100, 200)	-	-	-
Jul 2022	98 (95-100, 188)	100 (96-100, 102)	100 (98-100, 211)	-	-	-
Aug 2022	99 (96-100, 187)	100 (96-100, 100)	100 (98-100, 222)	-	-	-
Sep 2022	99 (96-100, 199)	100 (96-100, 100)	100 (98-100, 223)	-	84 (77-90, 102)	-
Oct 2022	98 (96-100, 251)	100 (97-100, 116)	100 (98-100, 223)	-	84 (77-90 102)	-
Nov 2022	98 (96-100, 255)	99 (95-100, 120)	100 (98-100, 224)	100 (99-100, 289)	84 (77-90, 102)	-
Dec 2022	98 (96-100, 255)	100 (97-100, 121)	100 (98-100, 226)	100 (99-100, 292)	84 (77-90, 102)	-
Jan 2023	98 (96-100, 255)	100 (97-100, 121)	100 (98-100, 225)	100 (99-100, 291)	-	-
Feb 2023	-	-	100 (98-100, 225)	100 (99-100, 292)	84 (76-90, 102)	-
Mar 2023	-	-	100 (98-100, 225)	100 (99-100, 292)	84 (76-90, 102)	-
Apr 2023	-	-	100 (98-100, 226)	100 (99-100, 292)	-	-
May 2023	-	-	100 (98-100, 194)	100 (99-100, 257)	84 (76-90, 102)	-
Jun 2023	-	-	100 (98-100, 225)	100 (99-100, 292)	-	-
Jul 2023	-	-	100 (98-100, 225)	100 (99-100, 292)	-	100 (97-100 112)
Aug 2023	-	-	100 (98-100, 226)	100 (99-100, 291)	-	-
Sep 2023	-	-	100 (98-100, 226)	100 (99-100, 291)	-	-

c) TNM stage

Data Cut	SYMPLIFY		RCRD % (95% CI, n)	NCRD % (95% CI, n)	DHCW % (95% CI, n)	WCISU % (95% CI, n)
	England % (95% CI, n)	Wales % (95% CI, n)				
Apr 2022	73 (65-79. 121)	71 (61-80, 62)	-	-	-	-
May 2022	72 (64-78, 133)	69 (59-78, 68)	-	-	-	-
Jun 2022	78 (71-83, 147)	72 (62-81, 73)	-	-	-	-
Jul 2022	79 (72-84, 150)	73 (63-81, 74)	-	-	-	-
Aug 2022	80 (73-85, 151)	74 (64-82, 74)	-	-	-	-
Sep 2022	81 (75-86, 163)	74 (64-82, 74)	-	-	-	-
Oct 2022	82 (77-87, 210)	72 (63-80, 84)	-	-	-	-
Nov 2022	83 (77-87, 214)	74 (65-81, 89)	-	74 (69-79, 214)	-	-
Dec 2022	83 (77-87, 214)	74 (65-81, 89)	-	74 (69-79, 216)	-	-
Jan 2023	83 (77-87, 214)	74 (65-81, 89)	-	74 (69-79, 216)	-	-
Feb 2023	-	-	-	75 (69-80, 218)	-	-
Mar 2023	-	-	-	76 (70-80, 221)	-	-
Apr 2023	-	-	-	76 (71-81, 222)	-	-
May 2023	-	-	-	74 (68-79, 190)	-	-
Jun 2023	-	-	-	76 (71-81, 223)	-	-
Jul 2023	-	-	-	76 (71-81, 223)	-	77 (68-84, 86)
Aug 2023	-	-	-	76 (71-81, 222)	-	-
Sep 2023	-	-	-	76 (71-81, 222)	-	-

d) Stage

Data Cut	SYMPLIFY		RCRD % (95% CI, n)	NCRD % (95% CI, n)	DHCW % (95% CI, n)	WCISU % (95% CI, n)
	England % (95% CI, n)	Wales % (95% CI, n)				
Apr 2022	99 (97-100, 165)	100 (96-100, 87)	71 (64-77, 140)	-	-	-
May 2022	99 (97-100, 185)	100 (96-100, 98)	70 (63-76, 136)	-	-	-
Jun 2022	99 (96-100, 187)	100 (96-100, 101)	70 (63-76, 139)	-	-	-
Jul 2022	99 (96-100, 189)	100 (96-100, 102)	71 (64-77, 150)	-	-	-
Aug 2022	99 (97-100, 188)	100 (96-100, 100)	71 (64-77, 157)	-	-	-
Sep 2022	100 (97-100, 201)	100 (96-100, 100)	70 (64-76, 157)	-	44 (35-53, 53)	-
Oct 2022	100 (99-100, 255)	100 (97-100, 116)	71 (64-77, 158)	-	44 (35-53, 53)	-
Nov 2022	100 (99-100, 259)	100 (97-100, 121)	71 (65-77, 160)	97 (94-98, 279)	44 (35-53, 53)	-
Dec 2022	100 (99-100, 259)	100 (97-100, 121)	72 (65-77, 162)	97 (94-98, 282)	44 (35-53, 53)	-
Jan 2023	100 (99-100, 259)	100 (97-100, 121)	72 (66-78, 162)	97 (94-99, 282)	-	-
Feb 2023	-	-	72 (66-78, 162)	98 (96-99, 287)	43 (34-53, 53)	-
Mar 2023	-	-	72 (66-78, 162)	100 (98-100, 291)	43 (34-53, 53)	-
Apr 2023	-	-	72 (65-77, 162)	100 (98-100, 291)	-	-
May 2023	-	-	68 (61-75, 132)	100 (98-100, 256)	43 (34-53, 53)	-
Jun 2023	-	-	72 (66-78, 162)	100 (98-100, 291)	-	-
Jul 2023	-	-	72 (66-78, 163)	100 (98-100, 290)	-	100 (97-100, 112)
Aug 2023	-	-	73 (66-78, 164)	100 (98-100, 290)	-	-
Sep 2023	-	-	73 (66-78, 164)	100 (98-100, 290)	-	-

There were 27, 55, 16, and 14 cancers reported in the central databases but not SYMPLIFY for RCRD, NCRD, DHCW, and WCISU, respectively. Table 3 demonstrates the breakdown of all cancers and whether they were reported in the central database, SYMPLIFY CRF, or both. Analysis of the cancers that were reported by NCRD but not SYMPLIFY-England at the last NCRD data cut available in September 2023, indicated that 55% (95% CI = 41%-68%, n=30) were diagnosed within the 3-month post-enrolment mandatory follow-up period in SYMPLIFY (Table 4). Similarly, there were 14 cancers that were reported by WCISU but not SYMPLIFY-Wales at the last WCISU data cut available in July 2023. Among these, 43% (95% CI = 18%-71%, n=6) were diagnosed within 3 months of enrolment (Table 4). There were 23 cancers reported in SYMPLIFY-England that were not reported in NCRD at the final time point available. Further investigations of these cancers showed that 61% (95% CI = 39%-80%, n=14) were, in fact, included in the registry data cuts but were excluded from the analysis due to the date of diagnosis preceding the date of study enrolment or due to an ineligible ICD-10 code reported by the registry (Table 5).

Colorectal cancers were the most frequently missed cancer sites in NCRD, RCRD, and SYMPLIFY-England, accounting for 35% (n=8), 33% (20), and 15% (8) of missed registrations, respectively (Table 6). The most common cancer site among missed cancers in DHCW, WCISU, and SYMPLIFY-Wales was lung, trachea, and bronchus, which accounted for 60% (9), 43% (10), and 21% (3) of missed registrations, respectively (Table 6). Among all cancer sites that were reported in SYMPLIFY-England, those with the highest percentage of missing cancers in NCRD were stomach (50%, n=3), cancer of unknown primary (50%, 1), and anus (33%, 1). In comparison, for RCRD, they were thyroid (100%, 1), bone and soft tissue (100%, 1), and pancreas (89%, n=8) (Table 7). Among all cancer sites that were reported in SYMPLIFY-Wales,

those with the highest percentage of missing cancers in WCISU were bone and soft tissue (males)/ovarian (females), breast, and cancer of unknown primary, whereby all cancers at these sites were missing in the registry (Table 7). As for DHCW, the cancer sites with the highest percentage of missing cancers were breast (100%, n=1) and ovarian (33%, 1) (Table 7).

Table 3. 2x2 tables demonstrating the breakdown of all cancers and whether they were reported in the central cancer database, SYMPLIFY CRF, or both for (a) RCRD, (b) NCRD, (c) DHCW, and (d) WCISU.

a) RCRD

		Central Database		**Total**
		YES	NO	
SYMPLIFY	YES	199	60	259
	NO	27	N/A	27
	Total	226	60	

b) NCRD

		Central Database		**Total**
		YES	NO	
SYMPLIFY	YES	235	23	259
	NO	55	N/A	55
	Total	291	23	

c) DHCW

		Central Database		**Total**
		YES	NO	
SYMPLIFY	YES	106	15	121
	NO	16	N/A	16
	Total	122	15	

d) WCISU

		Central Database		**Total**
		YES	NO	
SYMPLIFY	YES	98	23	121
	NO	14	N/A	14
	Total	112	23	

Table 4. Breakdown of the date of diagnosis relative to the date of study enrolment for cancers found in (a) NCRD (n=55) and (b) WCISU (n=14) that were not found in SYMPLIFY at the last registry data cut available (September 2023 and July 2023, respectively). Displayed as the number and percentage of missing cancers, and the cumulative number and percentage of missing cancers in each time frame.

a) NCRD – September 2023

Time Frame	Number of cancers diagnosed % (95% CI, n)	Cumulative diagnoses % (95% CI, n)
≤3 months post-enrolment	55 (41-68, 30)	55 (41-68, 30)
>3 and ≤6 months post-enrolment	31 (19-45, 17)	85 (73-94, 47)
>6 and ≤9 months post-enrolment	15 (6-27, 8)	100 (94-100, 55)

b) WCISU – July 2023

Time Frame	Number of cancers diagnosed % (95% CI, n)	Cumulative diagnoses % (95% CI, n)
≤3 months post-enrolment	43 (18-71, 6)	43 (18-71, 6)
>3 and ≤6 months post-enrolment	36 (13-65, 5)	79 (49-95, 11)
>6 and ≤9 months post-enrolment	21 (5-51, 3)	100 (77-100, 14)

Table 5. Investigation of cancers that were reported in SYMPLIFY-England but not reported in NCRD at the final time point in September 2023 (n=23).

Outcome of Investigation	Number of cancers % (95% CI, n)	Additional Notes
Cancer recorded in NCRD	61 (39-80, 14)	13 cancers were diagnosed pre-enrolment according to the date of diagnosis in NCRD 1 cancer in NCRD had an ineligible ICD-10 code resulting in its exclusion from the analysis
Cancer not recorded in NCRD	39 (20-61, 9)	6 cases involved NCRD reporting historical cancers, but no current cancer diagnoses related to those reported in SYMPLIFY-England 2 cases involved a patient with 3 cancers reported in SYMPLIFY-England but only 1 reported in NCRD 1 case involved no cancer diagnoses being reported in NCRD

Table 6. Cancer sites of cancers that were found in SYMPLIFY but not the respective national registry, and vice-versa. Displayed as the number and percentage of missed cancer registration cases at the last time point available for each dataset. For SYMPLIFY-England, displayed as the cancers found in the gold standard NCRD but not SYMPLIFY-England and for SYMPLIFY-Wales, displayed as the cancers found in the gold standard WCISU but not SYMPLIFY-Wales.

a) NCRD – September 2023 (n=23)

Cancer Site	**% (n)**
Colorectal	35 (8)
Stomach	13 (3)
Uterus	13 (3)
Bladder and urothelial	9 (2)
Oesophagus	9 (2)
Anus	4 (1)
Cancer of unknown primary	4 (1)
Lymphoid	4 (1)
Ovarian	4 (1)
Thyroid	4 (1)

b) RCRD – September 2023 (n=60)

Cancer Site	% (n)
Colorectal	33 (20)
Lung, trachea, bronchus	13 (8)
Pancreas	13 (8)
Bladder and urothelial	5 (3)
Oesophagus	5 (3)
Ovarian	5 (3)
Prostate	5 (3)
Stomach	5 (3)
Lymphoid	3 (2)
Other	3 (2)
Bone and soft tissue	2 (1)
Cancer of unknown primary	2 (1)
Liver, bile duct	2 (1)
Thyroid	2 (1)
Uterus	2 (1)

c) DHCW – May 2023 (n=15)

Cancer Site	% (n)
Lung, trachea, bronchus	60 (9)
Uterus	13 (2)
Breast, female	7 (1)
Other	7 (1)
Ovarian	7 (1)
Pancreas	7 (1)

e) WCISU – July 2023 (n=23)

Cancer Site	**% (n)**
Lung, trachea, bronchus	43 (10)
Bone and soft tissue (males) Ovarian (females)	9 (2)
Colorectal	9 (2)
Other	9 (2)
Ovarian	9 (2)
Uterus	9 (2)
Breast, female	4 (1)
Cancer of unknown primary	4 (1)
Pancreas	4 (1)

e) SYMPLIFY- England – January 2023 (n=55)

Cancer Site	**% (n)**
Colorectal	15 (8)
Lung, trachea, bronchus	13 (7)
Bladder and urothelial tract	11 (6)
Lymphoid	9 (5)
Oesophagus	7 (4)
Pancreas	7 (4)
Breast, female	5 (3)
Melanoma of skin	5 (3)
Other	5 (3)
Ovarian	5 (3)
Plasma cell	5 (3)
Liver, bile duct	4 (2)
Prostate	4 (2)
Head and neck	2 (1)
Uterus	2 (1)

f) SYMPLIFY-Wales – January 2023 (n=14)

Cancer Site	**% (n)**
Lung, trachea, bronchus	21 (3)
Colorectal	14 (2)
Prostate	14 (2)
Uterus	14 (2)
Breast, female	7 (1)
Lymphoid	7 (1)
Melanoma of skin	7 (1)
Other	7 (1)
Ovarian	7 (1)

Table 7. Proportion of missing cancers at each cancer site for (a) NCRD and RCRD, (b) DHCW and WCISU, (c) SYMPLIFY-England, and (d) SYMPLIFY-Wales. Displayed as the number and percentage of missed cancer registrations, based on the total number of cancers at that site reported in the comparator dataset.

a) NCRD and RCRD

SYMPLIFY – England (n=259) Cancer site (n)	NCRD Missing Cancers (n=23) % of total cancers at that site (n)	RCRD Missing Cancers (n=60) % of total cancers at that site (n)
Anus (3)	33 (1)	0 (0)
Bladder and urothelial tract (8)	25 (2)	38 (3)
Bone and soft tissue (1)	0 (0)	100 (1)
Breast, female (6)	0 (0)	0 (0)
Cancer of unknown primary (2)	50 (1)	50 (1)
Cervix (3)	0 (0)	0 (0)
Colorectal (121)	7 (8)	17 (20)
Gallbladder (1)	0 (0)	0 (0)
Head and neck (1)	0 (0)	0 (0)
Liver, bile duct (4)	0 (0)	25 (1)
Lung, trachea, bronchus (32)	0 (0)	25 (8)
Lymphoid (12)	8 (1)	17 (2)
Oesophagus (14)	14 (2)	21 (3)
Other (5)	0 (0)	40 (2)
Ovarian (10)	10 (1)	30 (3)
Pancreas (9)	0 (0)	89 (8)
Prostate (7)	0 (0)	43 (3)
Stomach (6)	50 (3)	50 (3)
Thyroid (1)	100 (1)	100 (1)
Uterus (13)	23 (3)	8 (1)

b) WCISU and DHCW

SYMPLIFY – Wales (n=121) Cancer site (n)	WCISU Missing Cancers (n=23) % of total cancers at that site (n)	DHCW Missing Cancers (n=15) % of total cancers at that site (n)
Anus (2)	0 (0)	0 (0)
Bladder and urothelial tract (2)	0 (0)	0 (0)
Bone and soft tissue (males), Ovarian (females) (2)	100 (2)	0 (0)
Breast, female (1)	100 (1)	100 (1)
CNS (1)	0 (0)	0 (0)
Cancer of unknown primary (1)	100 (1)	0 (0)
Cervix (1)	0 (0)	0 (0)
Colorectal (20)	10 (2)	0 (0)
Liver, bile duct (1)	0 (0)	0 (0)
Lung, trachea, bronchus (51)	20 (10)	18 (9)
Lymphoid (2)	0 (0)	0 (0)
Oesophagus (4)	0 (0)	0 (0)
Other (4)	50 (2)	25 (1)
Ovarian (3)	67 (2)	33 (1)
Pancreas (4)	25 (1)	25 (1)
Prostate (4)	0 (0)	0 (0)
Thyroid (1)	0 (0)	0 (0)
Uterus (17)	12 (2)	12 (2)

c) SYMPLIFY-England

NCRD (n=291) Cancer site (n)	SYMPLIFY-England Missing Cancers (n=55) % of total cancers at that site (n)
Anus (1)	0 (0)
Bladder and urothelial tract (12)	50 (6)
Bone and soft tissue (1)	0 (0)
Breast, female (9)	33 (3)
Cervix (3)	0 (0)
Colorectal (122)	7 (8)
Gallbladder (1)	0 (0)
Head and neck (2)	50 (1)
Liver, bile duct (7)	29 (2)
Lung, trachea, bronchus (39)	18 (7)
Lymphoid (17)	29 (5)
Melanoma of skin (3)	100 (3)
Oesophagus (19)	21 (4)
Other (8)	38 (3)
Ovarian (11)	27 (3)
Pancreas (13)	31 (4)
Plasma cell (3)	100 (3)
Prostate (8)	25 (2)
Uterus (12)	8 (1)

d) SYMPLIFY-Wales

WCISU (n=112) Cancer site (n)	SYMPLIFY-Wales Missing Cancers (n=14) % of total cancers at that site (n)
Anus (2)	0 (0)
Bladder and urothelial tract (2)	0 (0)
Breast, female (1)	100 (1)
CNS (1)	0 (0)
Cervix (1)	0 (0)
Colorectal (20)	10 (2)
Liver, bile duct (1)	0 (0)
Lung, trachea, bronchus (44)	7 (3)
Lymphoid (3)	33 (1)
Melanoma of skin (1)	100 (1)
Oesophagus (4)	0 (0)
Other (3)	33 (1)
Ovarian (2)	50 (1)
Pancreas (3)	0 (0)
Prostate (6)	33 (2)
Thyroid (1)	0 (0)
Uterus (17)	12 (2)

3.2. Concordance

Among the 259 cancers recorded in 250 participants in SYMPLIFY-England by September 2023, 199 cancers (77%, 95% CI = 71%-82%) in 197 participants were also recorded in RCRD, and 236 cancers (91%, 87%-94%) among 230 patients were also recorded in NCRD (Figure 4, Table 8). For SYMPLIFY-Wales, 121 cancers were reported among 118 patients by September 2023, of which 106 cancers (88%, 80%-93%) in 105 participants were also reported in DHCW, while 98 cancers (81%, 73%-88%) among 97 participants were also reported in WCISU (Figure 4, Table 8). The cancers reported in both SYMPLIFY and the respective central databases formed the population that was investigated for concordance.

Figure 4. Number of cancers identified in both SYMPLIFY and the corresponding central dataset over time.

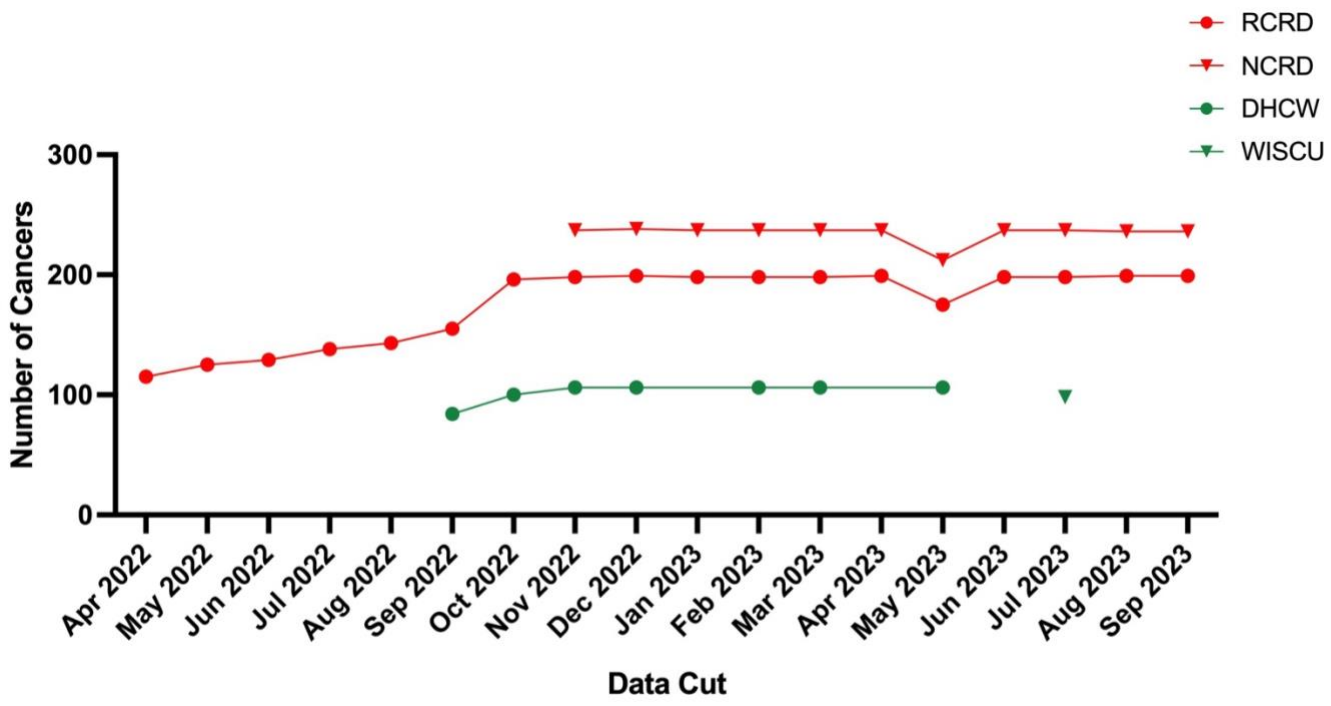


Table 8. Cancers found in both SYMPLIFY and the corresponding central dataset at each time point, which formed the population used in the concordance analysis. Proportions are derived from the denominator of 259 cancers recorded in the SYMPLIFY-England dataset for RCRD and NCRD datasets, and the 121 cancers recorded in the SYMPLIFY-Wales dataset for DHCW and WCISU.

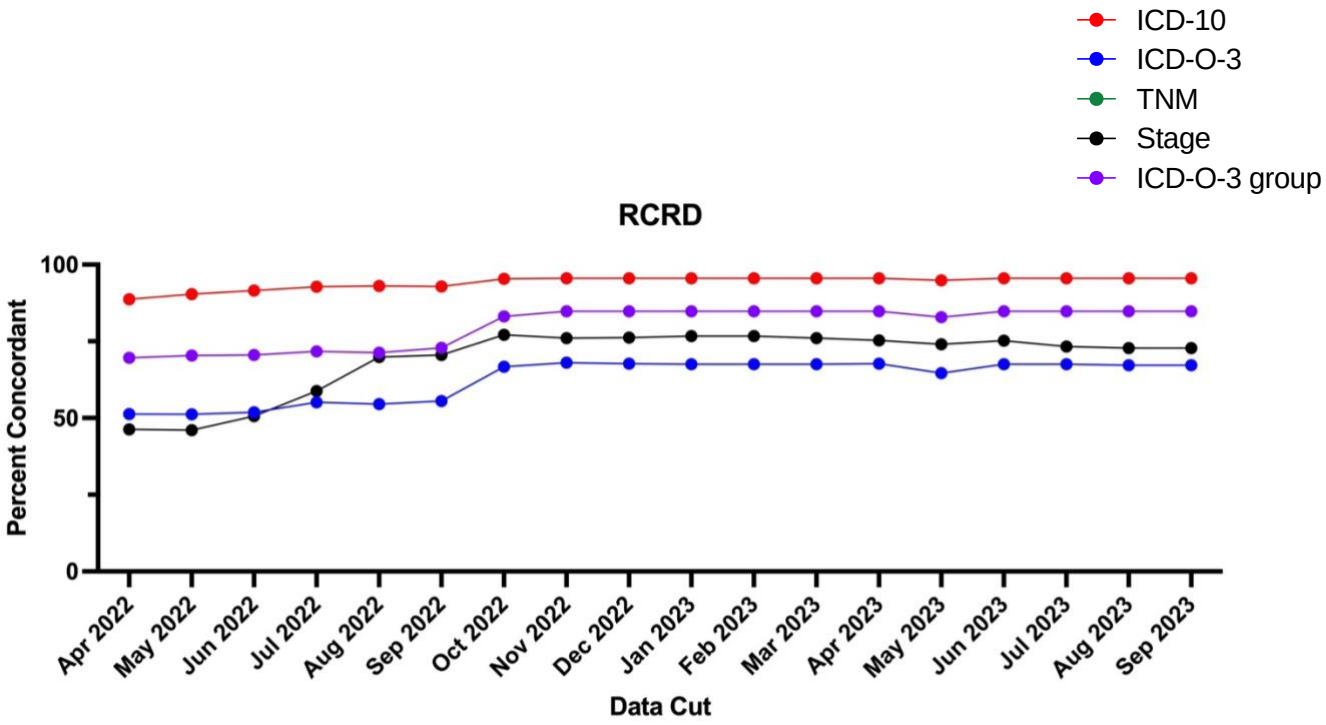
Data Cut	Dataset n cancers in dataset (n patients with cancer, % total SYMPLIFY cancers, 95% CI)			
	RCRD	**NCRD**	**DHCW**	**WCISU**
Apr 2022	115 (115, 44, 38-51)	-	-	-
May 2022	125 (125, 48, 42-55)	-	-	-
Jun 2022	129 (129, 50, 44-56)	-	-	-
Jul 2022	138 (138, 53, 47-59)	-	-	-
Aug 2022	143 (143, 55, 49-61)	-	-	-
Sep 2022	155 (155, 60, 54-66)	-	84 (83, 69, 60-77)	-
Oct 2022	196 (195, 76, 70-81)	-	100 (99, 83, 75-89)	-
Nov 2022	198 (197, 76, 71-81)	237 (231, 92, 87-95)	106 (105, 88, 80-93)	-
Dec 2022	199 (198, 77, 71-82)	238 (232, 92, 88-95)	106 (105, 88, 80-93)	-
Jan 2023	198 (197, 76, 71-81)	237 (231, 92, 87-95)	-	-
Feb 2023	198 (197, 76, 71-81)	237 (231, 92, 87-95)	106 (105, 88, 80-93)	-
Mar 2023	198 (197, 76, 71-81)	237 (231, 92, 87-95)	106 (105, 88, 80-93)	-
Apr 2023	199 (197, 77, 71-82)	237 (231, 92, 87-95)	-	-
May 2023	175 (173, 68, 61-73)	212 (206, 82, 77-86)	106 (105, 88, 80-93)	-
Jun 2023	198 (196, 76, 71-81)	237 (231, 92, 87-95)	-	-
Jul 2023	198 (196, 76, 71-81)	237 (231, 92, 87-95)	-	98 (97, 81, 73-88)
Aug 2023	199 (197, 77, 71-82)	236 (230, 91, 87-94)	-	-
Sep 2023	199 (197, 77, 71-82)	236 (230, 91, 87-94)	-	-

Concordance between SYMPLIFY-England and RCRD at the last time point available in September 2023 for ICD-10, ICD-O-3 morphology code, ICD-O-3 morphology groupings, and stage was 95% (95% CI = 92%-98%, n=190), 67% (60%-74%, 133), 85% (79%-90%, 168), and 73% (65%-80%, 107), respectively. Concordance between SYMPLIFY-England and NCRD in September 2023 varied by data point, whereby concordance was 96% (95% CI = 92%-98%, n=226), 60% (53%-66%, n=139), 83% (78%-88%, 193), 51% (44%-59%, 90), and 73% (67%-78%, 171), for ICD-10, ICD-O-3 code, ICD-O-3 morphology groupings, TNM, and stage, respectively (Figure 5, Table 9). Concordance between SYMPLIFY and NCRD was consistent over time, while RCRD demonstrated an increase in concordance for all data fields between April 2022 and October 2022 before remaining relatively consistent for the remainder of the study period (Figure 5, Table 9).

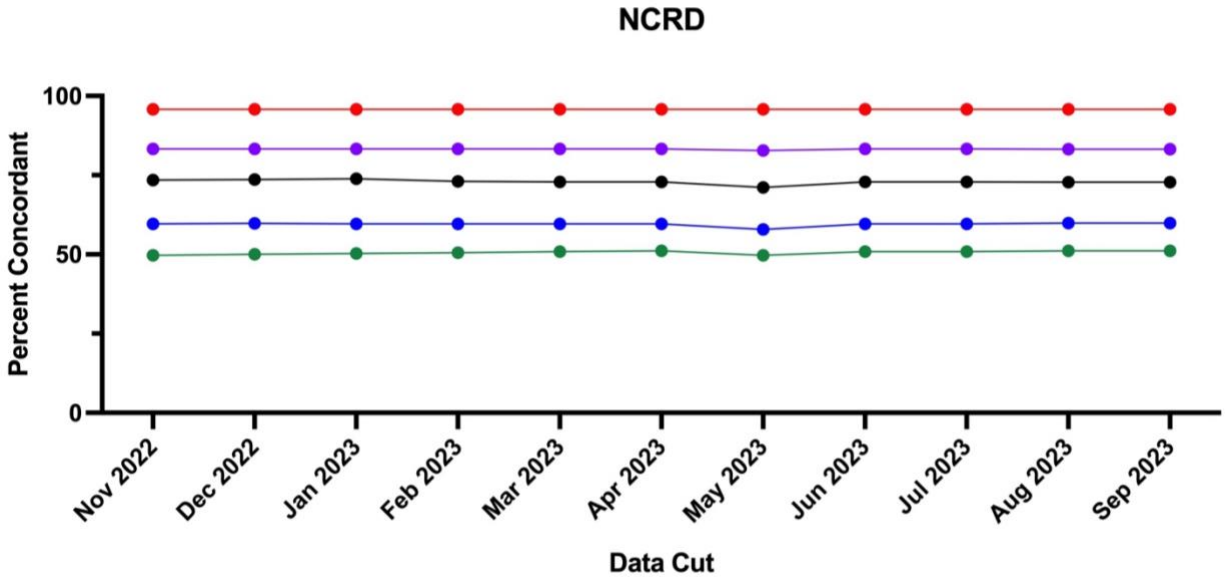
Concordance between SYMPLIFY-Wales and DHCW at the last time point available in May 2023 was 96% (95% CI = 91%-99%, n=102), 74% (64%-83%, 68), 84% (75%-91%, 77), and 87% (74%-95%, 40) for ICD-10, ICD-O-3 code, ICD-O-3 morphology groupings, and stage, respectively (Figure 5, Table 9). Concordance between SYMPLIFY-Wales and WCISU at the only time point available in July 2023 was 89% (95% CI = 81%-94%, n=87), 63% (53%-73%, 62), 80% (70%-87%, 78), 49% (38%-61%, 37), and 83% (74%-90%, 81) for ICD-10, ICD-O-3 code, ICD-O-3 morphology groupings, TNM, and stage, respectively (Figure 5, Table 9). There was a decrease in concordance in all data fields when the Welsh data switched from DHCW to WCISU. There was an increase in concordance based on the broader ICD-O-3 morphology groupings compared to the four-digit ICD-O-3 morphology code in all datasets (Figure 5, Table 9).

Figure 5. Concordance (%) between data fields in SYMPLIFY and panel (a) RCRD, (b) NCRD, and (c) DHCW and WCISU.

a)



b)



c)

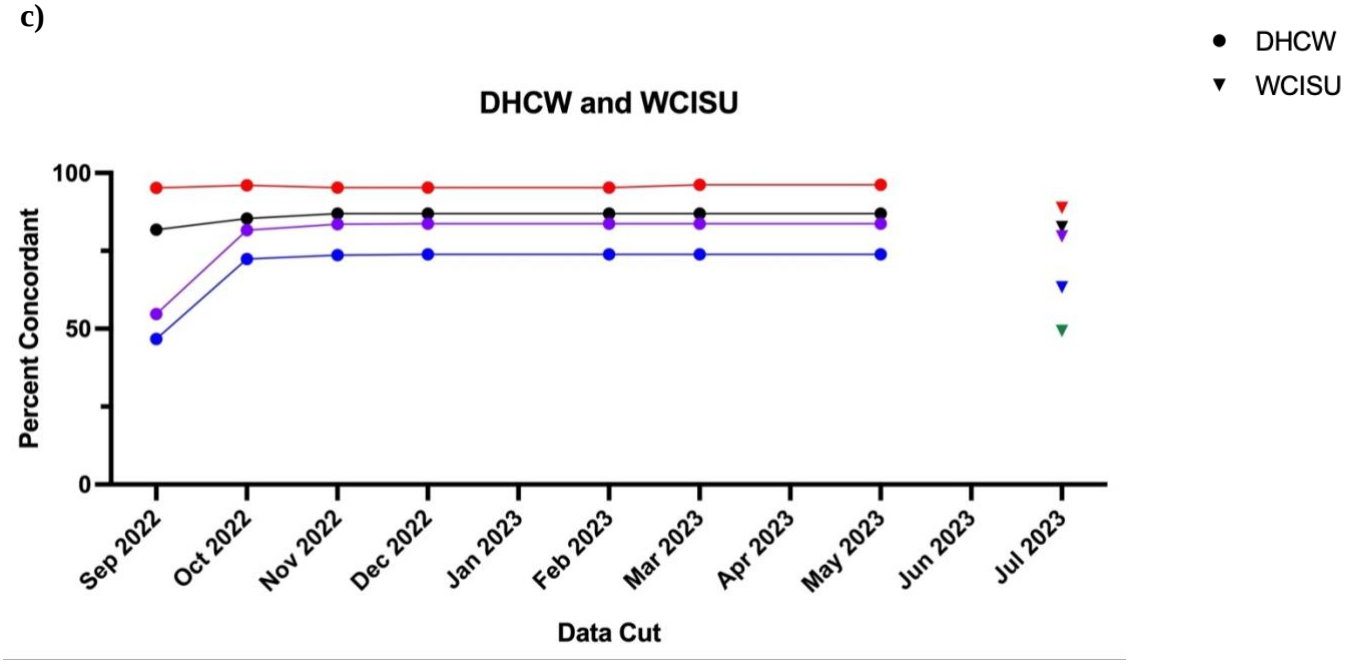

DHCW
WCISU
DHCW and WCISU
Percent Concordant
100
50
0
Sep 2022
Oct 2022
Nov 2022
Dec 2022
Jan 2023
Feb 2023
Mar 2023
Apr 2023
May 2023
Jun 2023
Jul 2023
Data Cut

Table 9. Concordance between cancer data fields in SYMPLIFY and the respective registries at each time point for (a) ICD-10, (b) ICD-O-3 4-digit morphology code, (c) ICD-O-3 broad morphology grouping, (d) TNM stage, and (e) stage. Displayed as the number and percentage of concordant cancers based on the total number of cancers reported in both SYMPLIFY and the respective registry with the data point completed in both datasets at that time point.

a) ICD-10

Data Cut	RCRD % (95% CI, n)	NCRD % (95% CI, n)	DHCW % (95% CI, n)	WCISU % (95% CI, n)
Apr 2022	89 (81-94, 102)	-	-	-
May 2022	90 (84-95, 113)	-	-	-
Jun 2022	91 (85-96, 118)	-	-	-
Jul 2022	93 (87-96, 128)	-	-	-
Aug 2022	93 (88-97, 133)	-	-	-
Sep 2022	93 (88-96, 144)	-	95 (88-99, 80)	-
Oct 2022	95 (91-98, 187)	-	96 (90-99, 98)	-
Nov 2022	95 (92-98, 189)	96 (92-98, 227)	95 (89-98, 101)	-
Dec 2022	95 (92-98, 190)	96 (92-98, 228)	95 (89-98, 101)	-
Jan 2023	95 (92-98, 189)	96 (92-98, 227)	-	-
Feb 2023	95 (92-98, 189)	96 (92-98, 227)	95 (89-98, 101)	-
Mar 2023	95 (92-98, 189)	96 (92-98, 227)	96 (89-98, 102)	-
Apr 2023	95 (92-98, 190)	96 (92-98, 227)	-	-
May 2023	95 (90-98, 166)	96 (92-98, 203)	96 (91-99, 102)	-
Jun 2023	95 (92-98, 189)	96 (92-98, 227)	-	-
Jul 2023	95 (92-98, 189)	96 (92-98, 227)	-	89 (81-94, 87)
Aug 2023	95 (92-98, 190)	96 (92-98, 226)	-	-
Sep 2023	95 (92-98, 190)	96 (92-98, 226)	-	-

b) ICD-O-3 4-digit morphology code

Data Cut	RCRD % (95% CI, n)	NCRD % (95% CI, n)	DHCW % (95% CI, n)	WCISU % (95% CI, n)
Apr 2022	51 (42-61, 59)	-	-	-
May 2022	51 (42-60, 64)	-	-	-
Jun 2022	52 (43-61, 67)	-	-	-
Jul 2022	55 (46-64, 76)	-	-	-
Aug 2022	55 (46-63, 78)	-	-	-
Sep 2022	55 (47-63, 86)	-	47 (35-59, 35)	-
Oct 2022	67 (60-73, 130)	-	72 (62-81, 63)	-
Nov 2022	68 (61-74, 134)	60 (53-66, 139)	74 (63-82, 67)	-
Dec 2022	68 (61-74, 134)	60 (53-66, 140)	74 (64-83, 68)	-
Jan 2023	68 (60-74, 133)	60 (53-66, 139)	-	-
Feb 2023	68 (60-74, 133)	60 (53-66, 139)	74 (64-83, 68)	-
Mar 2023	68 (60-74, 133)	60 (53-66, 139)	74 (64-83, 68)	-
Apr 2023	68 (61-74, 134)	60 (53-66, 139)	-	-
May 2023	65 (57-72, 113)	58 (51-65, 121)	74 (64-83, 68)	-
Jun 2023	68 (60-74, 133)	60 (53-66, 139)	-	-
Jul 2023	68 (60-74, 133)	60 (53-66, 139)	-	63 (53-73, 62)
Aug 2023	67 (60-74, 133)	60 (53-66, 139)	-	-
Sep 2023	67 (60-74, 133)	60 (53-66, 139)	-	-

c) ICD-O-3 broad morphology grouping

Data Cut	RCRD % (95% CI, n)	NCRD % (95% CI, n)	DHCW % (95% CI, n)	WCISU % (95% CI, n)
Apr 2022	70 (60-78, 80)	-	-	-
May 2022	70 (62-78, 88)	-	-	-
Jun 2022	71 (62-78, 91)	-	-	-
Jul 2022	72 (63-79, 99)	-	-	-
Aug 2022	71 (63-79, 102)	-	-	-
Sep 2022	73 (65-78, 113)	-	55 (43-66, 41)	-
Oct 2022	83 (77-88, 162)	-	82 (72-89, 71)	-
Nov 2022	85 (79-89, 167)	83 (78-88, 194)	84 (74-90, 76)	-
Dec 2022	85 (79-90, 168)	83 (78-88, 195)	84 (75-91, 77)	-
Jan 2023	85 (79-89, 167)	83 (78-88, 194)	-	-
Feb 2023	85 (79-89, 167)	83 (78-88,194)	84 (75-91, 77)	-
Mar 2023	85 (79-89, 167)	83 (78-88, 194)	84 (75-91, 77)	-
Apr 2023	85 (79-90, 168)	83 (78-88, 194)	-	-
May 2023	83 (76-88, 145)	83 (77-88, 173)	84 (75-91, 77)	-
Jun 2023	85 (79-89, 167)	83 (78-88, 194)	-	-
Jul 2023	85 (79-89, 167)	83 (78-88, 194)	-	80 (70-87, 78)
Aug 2023	85 (79-90, 168)	83 (78-88, 193)	-	-
Sep 2023	85 (79-90, 168)	83 (78-88, 193)	-	-

d) TNM stage

Data Cut	RCRD % (95% CI, n)	NCRD % (95% CI, n)	DHCW % (95% CI, n)	WCISU % (95% CI, n)
Apr 2022	-	-	-	-
May 2022	-	-	-	-
Jun 2022	-	-	-	-
Jul 2022	-	-	-	-
Aug 2022	-	-	-	-
Sep 2022	-	-	-	-
Oct 2022	-	-	-	-
Nov 2022	-	50 (42-57, 86)	-	-
Dec 2022	-	50 (42-58, 87)	-	-
Jan 2023	-	50 (43-58, 87)	-	-
Feb 2023	-	51 (43-58, 88)	-	-
Mar 2023	-	51 (43-58, 89)	-	-
Apr 2023	-	51 (44-59, 90)	-	-
May 2023	-	50 (41-58, 76)	-	-
Jun 2023	-	51 (43-58, 90)	-	-
Jul 2023	-	51 (43-58, 90)	-	49 (38-61, 37)
Aug 2023	-	51 (44-59, 90)	-	-
Sep 2023	-	51 (44-59, 90)	-	-

e) Stage

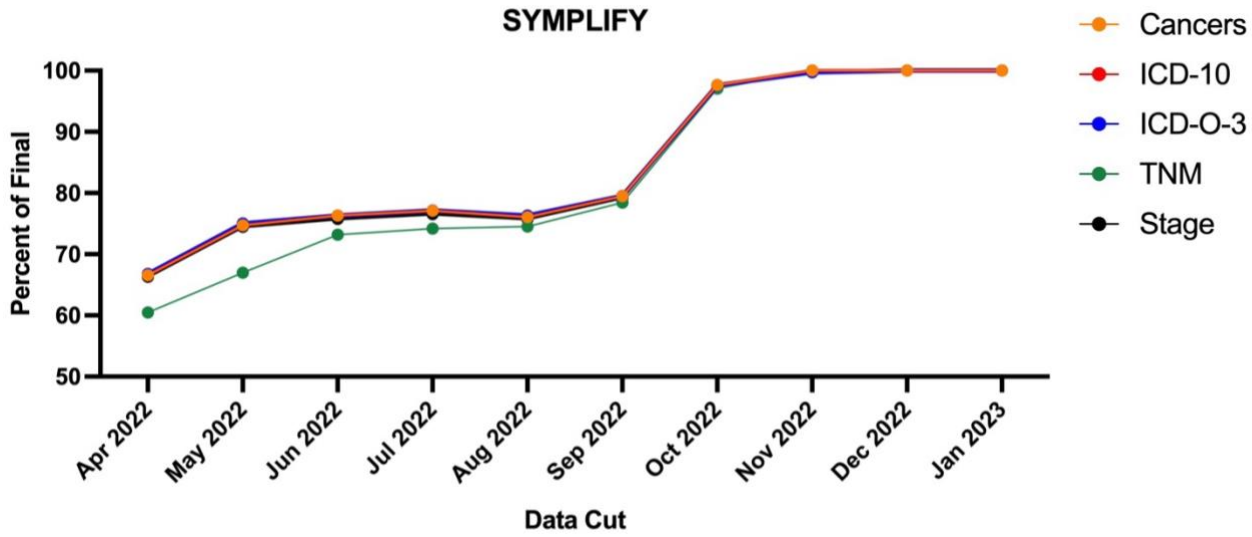
Data Cut	RCRD % (95% CI, n)	NCRD % (95% CI, n)	DHCW % (95% CI, n)	WCISU % (95% CI, n)
Apr 2022	46 (35-58, 38)	-	-	-
May 2022	46 (35-57, 40)	-	-	-
Jun 2022	51 (40-61, 45)	-	-	-
Jul 2022	59 (48-69, 57)	-	-	-
Aug 2022	70 (60-79, 72)	-	-	-
Sep 2022	71 (61-79, 79)	-	82 (65-93, 27)	-
Oct 2022	77 (69-84, 111)	-	85 (71-94, 35)	-
Nov 2022	76 (68-83, 111)	73 (67-79, 169)	87 (74-95, 40)	-
Dec 2022	76 (68-83, 112)	74 (67-79, 170)	87 (74-95, 40)	-
Jan 2023	77 (69-83, 112)	74 (68-79, 170)	-	-
Feb 2023	77 (69-83, 112)	73 (67-79, 171)	87 (74-95, 40)	-
Mar 2023	76 (68-83, 111)	73 (67-78, 172)	87 (74-95, 40)	-
Apr 2023	75 (68-82, 110)	73 (67-78, 172)	-	-
May 2023	74 (65-81, 91)	71 (64-77, 150)	87 (74-95, 40)	-
Jun 2023	75 (67-82, 109)	73 (67-78, 172)	-	-
Jul 2023	73 (65-80, 107)	73 (67-78, 172)	-	83 (74-90, 81)
Aug 2023	73 (65-80, 107)	73 (67-78, 171)	-	-
Sep 2023	73 (65-80, 107)	73 (67-78, 171)	-	-

3.3. Timeliness

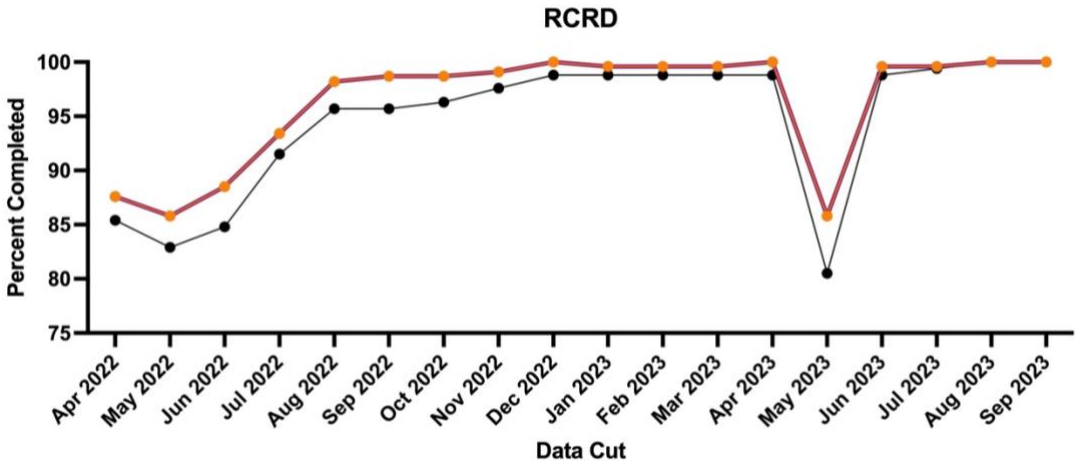
SYMPLIFY reached completion of cancer registrations in November 2022, approximately 12 months after the study recruitment ended on November 30, 2021 (Figure 6, Table 10). This dataset also reached full concordance with the final most up-to-date data cut at 12 months post-enrolment for ICD-10, TNM stage, and overall stage and at 13 months post-enrolment for ICD-O-3 (Figure 7, Table 11). NCRD and RCRD both reached completion of cancer registrations compared to the corresponding final data cuts in December 2022, at 13 months post-enrolment (Figure 6, Table 10). Concordance with the final data cut was over 90% for all data points at 12 months post-enrolment for NCRD. However, complete concordance was reached at 13, 16, 17, and 21 months post-enrolment for ICD-10, stage, TNM classification, and ICD-O-3, respectively (Figure 7, Table 11). Similarly, the DHCW dataset reached full completeness of cancer registration at 15 months post-enrolment, while complete concordance was reached at 10 months post-enrolment for ICD-O-3 and stage and at 16 months post-enrolment for ICD-10.

Figure 6. Timeliness of the completeness of data fields for panel (a) SYMPLIFY, (b) RCRD, (c) NCRD, and (d) DHCW compared to the final dataset for each data source. Displayed as the proportion of cancers with the completed data field compared to the total number of cancers with the completed data field in the final data cut. Cancer number, ICD-10, ICD-O-3, and stage overlap in SYMPLIFY, RCRD, and NCRD. Cancer number and ICD-10 overlap in DHCW, as do ICD-O-3 and stage.

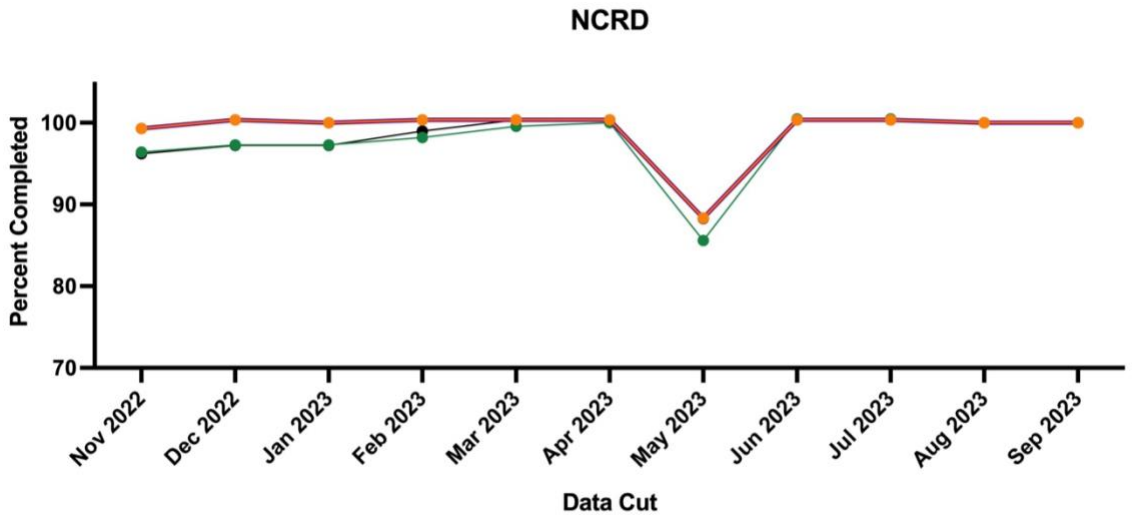
a)



b)



c)


NCRD
Percent Completed
100
90
80
70
Nov 2022
Dec 2022
Jan 2023
Feb 2023
Mar 2023
Apr 2023
May 2023
Jun 2023
Jul 2023
Aug 2023
Sep 2023
Data Cut

d)

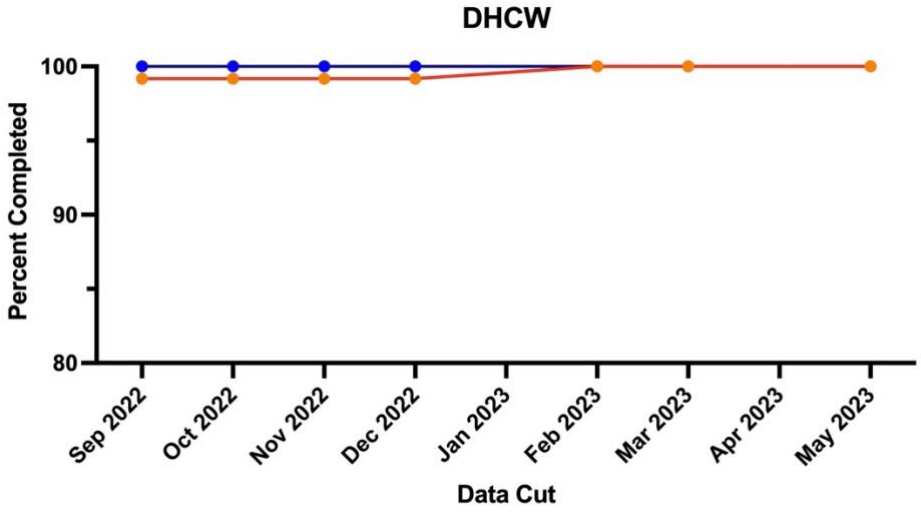

DHCW
Percent Completed
100
90
80
Sep 2022
Oct 2022
Nov 2022
Dec 2022
Jan 2023
Feb 2023
Mar 2023
Apr 2023
May 2023
Data Cut

Table 10. Timeliness of the completeness of each data field for (a) SYMPLIFY, (b) RCRD, (c) NCRD, and (d) DHCW datasets. Displayed as the number and proportion of the total number of cancers with the completed data field in the final data cut for each dataset.

a) SYMPLIFY

Data Cut	Data Field % (95% CI, n)				
	Cancers	**ICD-10**	**ICD-O-3**	**TNM stage**	**Stage**
Apr 2022	67 (62-71, 253)	67 (62-71, 253)	67 (62-72, 251)	60 (55-66, 185)	66 (61-71, 252)
May 2022	75 (70-79, 284)	75 (70-79, 284)	75 (70-79, 282)	67 (61-72, 205)	74 (70-79, 283)
Jun 2022	76 (72-81, 290)	76 (72-81, 290)	76 (72-81, 287)	73 (68-78, 224)	76 (71-80, 288)
Jul 2022	77 (73-81, 293)	77 (73-81, 293)	77 (73-81, 290)	74 (69-79, 227)	77 (72-81, 291)
Aug 2022	76 (71-80, 289)	76 (71-80, 289)	76 (72-81, 287)	75 (69-79, 228)	76 (71-80, 288)
Sep 2022	79 (75-83, 302)	79 (75-83, 302)	80 (75-83, 299)	78 (73-83, 240)	79 (75-83, 301)
Oct 2022	98 (96-99, 371)	98 (96-99, 371)	98 (96-99, 367)	97 (94-99, 297)	98 (96-99, 371)
Nov 2022	100 (99-100, 380)	100 (99-100, 380)	100 (99-100, 375)	100 (99-100, 306)	100 (99-100, 380)
Dec 2022	100 (99-100, 380)	100 (99-100, 380)	100 (99-100, 376)	100 (99-100, 306)	100 (99-100, 380)
Jan 2023	100 (99-100, 380)	100 (99-100, 380)	100 (99-100, 376)	100 (99-100, 306)	100 (99-100, 380)

b) RCRD

Data Cut	Data Field % (95% CI, n)				
	Cancers	**ICD-10**	**ICD-O-3**	**TNM stage**	**Stage**
Apr 2022	88 (83-92, 198)	88 (83-92, 198)	88 (83-92, 198)	-	85 (79-90, 140)
May 2022	86 (81-90, 194)	86 (81-90, 194)	86 (81-90, 194)	-	83 (76-88, 136)
Jun 2022	88 (84-92, 200)	88 (84-92, 200)	88 (84-92, 200)	-	85 (78-90, 139)
Jul 2022	93 (89-96, 211)	93 (89-96, 211)	93 (89-96, 211)	-	91 (86-95, 150)
Aug 2022	98 (96-100, 222)	98 (96-100, 222)	98 (96-100, 222)	-	96 (91-98, 157)
Sep 2022	99 (96-100, 223)	99 (96-100, 223)	99 (96-100, 223)	-	96 (91-98, 157)
Oct 2022	99 (96-100, 223)	99 (96-100, 223)	99 (96-100, 223)	-	96 (92-99, 158)
Nov 2022	99 (97-100, 224)	99 (97-100, 224)	99 (97-100, 224)	-	98 (94-99, 160)
Dec 2022	100 (98-100, 226)	100 (98-100, 226)	100 (98-100, 226)	-	99 (96-100, 162)
Jan 2023	100 (98-100, 225)	100 (98-100, 225)	100 (98-100, 225)	-	99 (96-100, 162)
Feb 2023	100 (98-100, 225)	100 (98-100, 225)	100 (98-100, 225)	-	99 (96-100, 162)
Mar 2023	100 (98-100, 225)	100 (98-100, 225)	100 (98-100, 225)	-	99 (96-100, 162)
Apr 2023	100 (98-100, 226)	100 (98-100, 226)	100 (98-100, 226)	-	99 (96-100, 162)
May 2023	86 (81-90, 194)	86 (81-90, 194)	86 (81-90, 194)	-	80 (74-86, 132)
Jun 2023	100 (98-100, 225)	100 (98-100, 225)	100 (98-100, 225)	-	99 (96-100, 162)
Jul 2023	100 (98-100, 225)	100 (98-100, 225)	100 (98-100, 225)	-	99 (97-100, 163)
Aug 2023	100 (98-100, 226)	100 (98-100, 226)	100 (98-100, 226)	-	100 (98-100, 164)
Sep 2023	100 (98-100, 226)	100 (98-100, 226)	100 (98-100, 226)	-	100 (98-100, 164)

c) NCRD

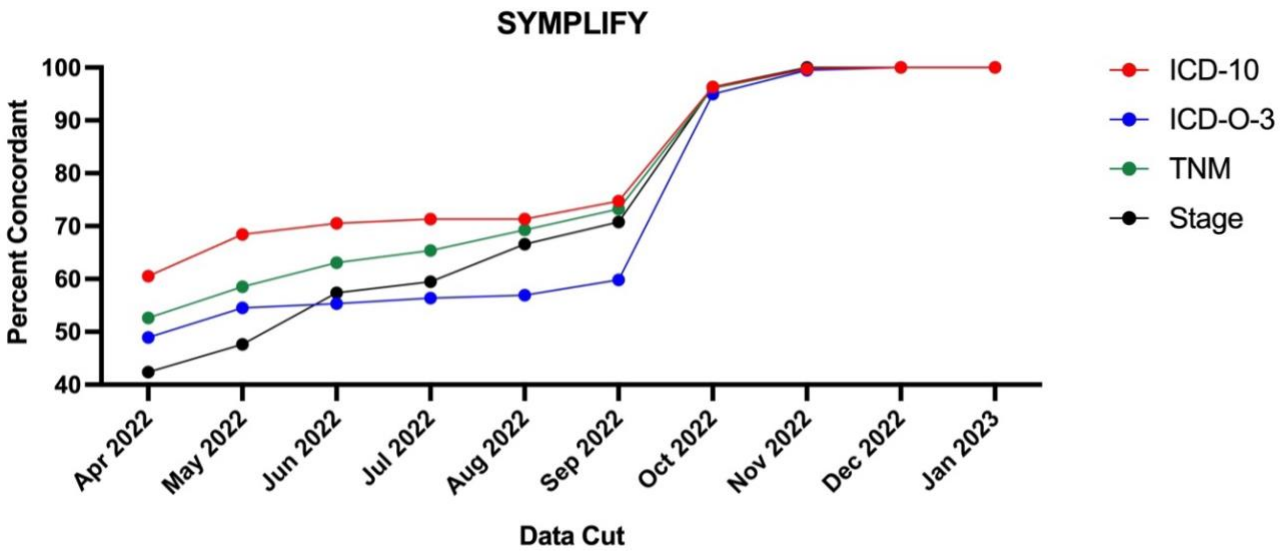
Data Cut	Data Field % (95% CI, n)				
	Cancers	**ICD-10**	**ICD-O-3**	**TNM stage**	**Stage**
Nov 2022	99 (98-100, 289)	99 (98-100, 289)	99 (98-100, 289)	96 (93-98, 214)	96 (93-98, 279)
Dec 2022	100 (99-100, 292)	100 (99-100, 292)	100 (99-100, 292)	97 (94-99, 216)	97 (95-99, 282)
Jan 2023	100 (99-100, 291)	100 (99-100, 291)	100 (99-100, 291)	97 (94-99, 216)	97 (95-99, 282)
Feb 2023	100 (99-100, 292)	100 (99-100, 292)	100 (99-100, 292)	98 (95-100, 218)	99 (97-100, 287)
Mar 2023	100 (99-100, 292)	100 (99-100, 292)	100 (99-100, 292)	100 (98-100, 221)	100 (99-100, 291)
Apr 2023	100 (99-100, 292)	100 (99-100, 292)	100 (99-100, 292)	100 (98-100, 222)	100 (99-100, 291)
May 2023	88 (84-92, 257)	88 (84-92, 257)	88 (84-92, 257)	86 (80-90, 190)	88 (84-92, 256)
Jun 2023	100 (99-100, 292)	100 (99-100, 292)	100 (99-100, 292)	100 (98-100, 223)	100 (99-100, 291)
Jul 2023	100 (99-100, 292)	100 (99-100, 292)	100 (99-100, 292)	100 (98-100, 223)	100 (99-100, 291)
Aug 2023	100 (99-100, 291)	100 (99-100, 291)	100 (99-100, 291)	100 (98-100, 222)	100 (99-100, 290)
Sep 2023	100 (99-100, 291)	100 (99-100, 291)	100 (99-100, 291)	100 (98-100, 222)	100 (99-100, 290)

d) DHCW

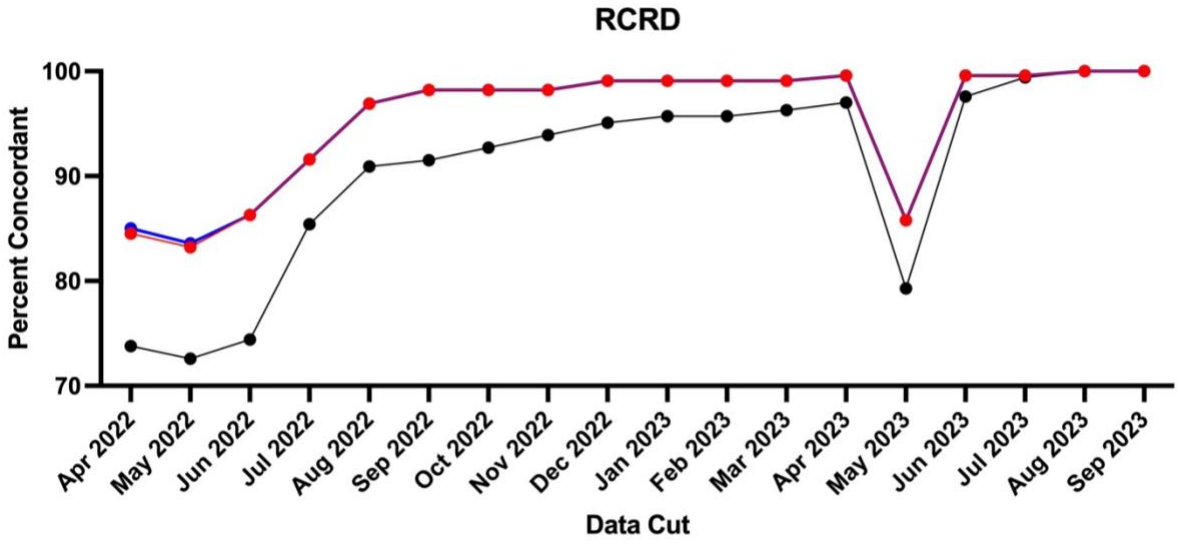
Data Cut	Data Field % (95% CI, n)				
	Cancers	**ICD-10**	**ICD-O-3**	**TNM stage**	**Stage**
Sep 2022	99 (96-100, 121)	99 (96-100, 121)	100 (96-100, 102)	-	100 (93-100, 53)
Oct 2022	99 (96-100, 121)	99 (96-100, 121)	100 (96-100, 102)	-	100 (93-100, 53
Nov 2022	99 (96-100, 121)	99 (96-100, 121)	100 (96-100, 102)	-	100 (93-100, 53
Dec 2022	99 (96-100, 121)	99 (96-100, 121)	100 (96-100, 102)	-	100 (93-100, 53
Jan 2023	-	-	-	-	-
Feb 2023	100 (97-100, 122)	100 (97-100, 122)	100 (96-100, 102)	-	100 (93-100, 53
Mar 2023	100 (97-100, 122)	100 (97-100, 122)	100 (96-100, 102)	-	
Apr 2023	-	-	-	-	-
May 2023	100 (97-100, 122)	100 (97-100, 122)	100 (96-100, 102)	-	100 (93-100, 53

Figure 7. Timeliness of the concordance of data fields compared to the final data cut for (a) SYMPLIFY, (b) RCRD, (c), NCRD, and (d) DHCW. Calculated as the proportion (%) of concordant cancers based on the total number of cancers with the completed data field in the final data cut. ICD-10 and ICD-O-3 overlap in RCRD, while ICD-O-3 and stage overlap in DHCW.

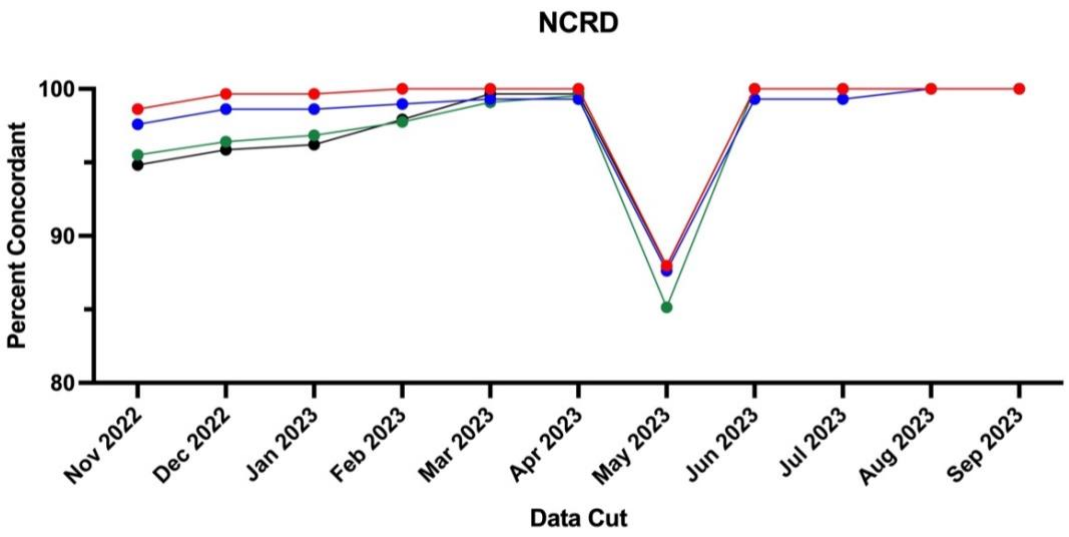
a)



b)



c)


NCRD
Percent Concordant
100
90
80
Nov 2022
Dec 2022
Jan 2023
Feb 2023
Mar 2023
Apr 2023
May 2023
Jun 2023
Jul 2023
Aug 2023
Sep 2023
Data Cut

d)

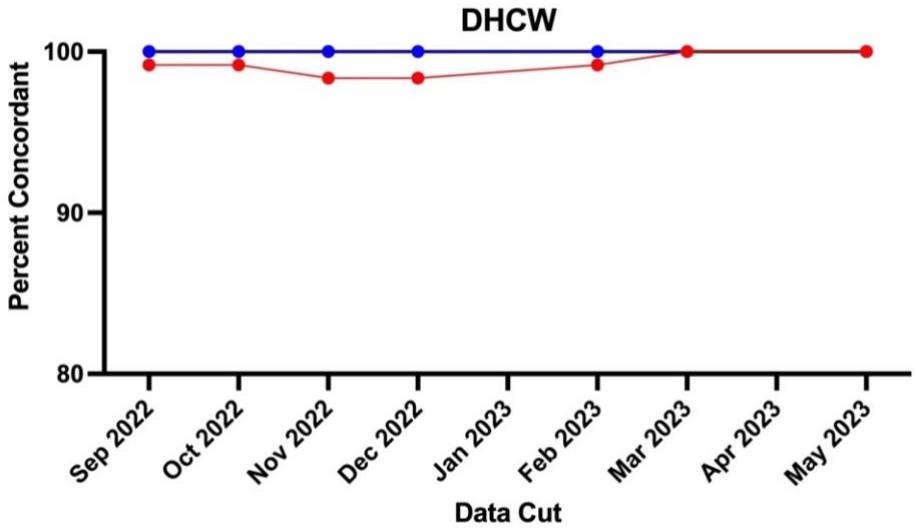

DHCW
Percent Concordant
100
90
80
Sep 2022
Oct 2022
Nov 2022
Dec 2022
Jan 2023
Feb 2023
Mar 2023
Apr 2023
May 2023
Data Cut

Table 11. Timeliness of the concordance of each data field for (a) SYMPLIFY, (b) RCRD, (c) NCRD, and (d) DHCW datasets. Presented as the number and proportion of concordant cancers based on the total number of cancers with the completed data field in the final data cut.

a) SYMPLIFY

Data Cut	Data Field % (95% CI, n)			
	ICD-10	**ICD-O-3**	**TNM stage**	**Stage**
Apr 2022	61 (55-65, 230)	49 (44-54, 184)	53 (47-58, 161)	42 (37-48, 161)
May 2022	68 (63-73, 260)	55 (49-60, 205)	58 (53-64, 179)	48 (43-53, 181)
Jun 2022	71 (66-75, 268)	55 (50-60, 208)	63 (57-68, 193)	57 (52-62, 218)
Jul 2022	71 (66-76, 271)	56 (51-61, 212)	65 (60-71, 200)	59 (54-64, 226)
Aug 2022	71 (66-76, 271)	57 (52-62, 214)	69 (64-74, 212)	67 (62-71, 253)
Sep 2022	75 (70-79, 284)	60 (55-65, 225)	73 (68-78, 224)	71 (66-75, 269)
Oct 2022	96 (94-98, 366)	95 (92-97, 357)	96 (93-98, 294)	96 (94-98, 366)
Nov 2022	100 (99-100, 379)	99 (98-100, 374)	100 (98-100, 305)	100 (99-100, 380)
Dec 2022	100 (99-100, 380)	100 (99-100, 376)	100 (99-100, 306)	100 (99-100, 380)
Jan 2023	100 (99-100, 380)	100 (99-100, 376)	100 (99-100, 306)	100 (99-100, 380)

b) RCRD

Data Cut	Data Field % (95% CI, n)			
	ICD-10	**ICD-O-3**	**TNM stage**	**Stage**
Apr 2022	85 (79-89, 191)	85 (80-89, 192)	-	74 (66-80, 121)
May 2022	83 (78-88, 188)	84 (78-88, 189)	-	73 (65-79, 119)
Jun 2022	86 (81-90, 195)	86 (81-90, 195)	-	74 (67-81, 122)
Jul 2022	92 (87-95, 207)	92 (87-95, 207)	-	85 (79-90, 140)
Aug 2022	97 (94-99, 219)	97 (94-99, 219)	-	91 (85-95, 149)
Sep 2022	98 (96-100, 222)	98 (96-100, 222)	-	91 (86-95, 150)
Oct 2022	98 (96-100, 222)	98 (96-100, 222)	-	93 (88-96, 152)
Nov 2022	98 (96-100, 222)	98 (96-100, 222)	-	94 (89-97, 154)
Dec 2022	99 (97-100, 224)	99 (97-100, 224)	-	95 (91-98, 156)
Jan 2023	99 (97-100, 224)	99 (97-100, 224)	-	96 (91-98, 157)
Feb 2023	99 (97-100, 224)	99 (97-100, 224)	-	96 (91-98, 157)
Mar 2023	99 (97-100, 224)	99 (97-100, 224)	-	96 (92-99, 158)
Apr 2023	99 (98-100, 224)	100 (98-100, 225)	-	97 (93-99, 159)
May 2023	86 (81-90, 194)	86 (81-90, 194)	-	79 (72-85, 130)
Jun 2023	100 (98-100, 225)	100 (98-100, 225)	-	98 (94-99, 160)
Jul 2023	100 (98-100, 225)	100 (98-100, 225)	-	99 (97-100, 163)
Aug 2023	100 (98-100, 226)	100 (98-100, 226)	-	100 (98-100, 164)
Sep 2023	100 (98-100, 226)	100 (98-100, 226)	-	100 (98-100, 164)

c) NCRD

Data Cut	Data Field % (95% CI, n)			
	ICD-10	**ICD-O-3**	**TNM stage**	**Stage**
Nov 2022	99 (97-100, 287)	98 (95-99, 284)	95 (92-98, 212)	95 (92-97, 275)
Dec 2022	100 (98-100, 290)	99 (97-100, 287)	96 (93-98, 214)	96 (93-98, 278)
Jan 2023	100 (98-100, 290)	99 (97-100, 287)	97 (94-99, 215)	96 (93-98, 279)
Feb 2023	100 (99-100, 291)	99 (97-100, 288)	98 (95-99, 217)	98 (96-99, 284)
Mar 2023	100 (99-100, 291)	99 (98-100, 289)	99 (97-100, 220)	100 (98-100, 289)
Apr 2023	100 (99-100, 291)	99 (98-100, 289)	100 (98-100, 221)	100 (98-100, 289)
May 2023	88 (84-91, 256)	88 (83-91, 255)	85 (80-90, 189)	88 (84-91, 255)
Jun 2023	100 (99-100, 291)	99 (98-100, 289)	100 (98-100, 222)	100 (99-100, 290)
Jul 2023	100 (99-100, 291)	99 (98-100, 289)	100 (98-100, 222)	100 (99-100, 290)
Aug 2023	100 (99-100, 291)	100 (99-100, 291)	100 (98-100, 222)	100 (99-100, 290)
Sep 2023	100 (99-100, 291)	100 (99-100, 291)	100 (98-100, 222)	100 (99-100, 290)

d) DHCW

Data Cut	Data Field % (95% CI, n)		
	ICD-10	**ICD-O-3**	**Stage**
Sep 2022	99 (96-100, 121)	100 (96-100, 102)	100 (93-100, 53)
Oct 2022	99 (96-100, 121)	100 (96-100, 102)	100 (93-100, 53)
Nov 2022	98 (94-100, 120)	100 (96-100, 102)	100 (93-100, 53)
Dec 2022	98 (94-100, 120)	100 (96-100, 102)	100 (93-100, 53)
Jan 2023	-	-	-
Feb 2023	99 (96-100, 121)	100 (96-100, 102)	100 (93-100, 53)
Mar 2023	100 (97-100, 122)	100 (96-100, 102)	100 (93-100, 53)
Apr 2023	-	-	-
May 2023	100 (97-100, 122)	100 (96-100, 102)	100 (93-100, 53)

3.4. Completeness by Cancer Site and Referral Pathway

SYMPLIFY demonstrated low completeness of TNM staging for lymphoid (14%, 95% CI = 2%-43%), ovarian (38%, 14%-68%), and uterine (20%, 8%-39%) cancers but high completeness for oesophageal (100%, 81%-100%), prostate (100%, 72%-100%), and colorectal (95%, 90%-98%) cancers (Table 12). NCRD also demonstrated similarly low TNM completeness for lymphoid (0%, 0%-20%) and ovarian (27%, 6%-61%) cancers, while both NCRD and WCISU had low completeness of uterine cancers at 25% (5%-57%) and 24% (7%-50%), respectively. DHCW demonstrated a low stage completeness of lung cancers (4%, 1%-15%) compared to the other datasets, which demonstrated stage completeness of over 95% for the lung cancers reported (Table 12). Regarding cancer referral pathways, stage completeness was lowest for cancers investigated via the gynae 2WW pathway and rapid diagnostic centres for RCRD, while cancers investigated by the lung 2WW pathway demonstrated the lowest stage completeness (12%, 4%-23%) in DHCW (Table 12).

Table 12. Completeness of data fields based on cancer site and diagnostic referral pathway in the final data cut for (a) SYMPLIFY, (b) RCRD, (c) NCRD, (d) DHCW, and (e) WCISU. Displayed as the number and percentage of total cancers at the given cancer site or diagnosed via the referral pathway.

a) SYMPLIFY

SYMPLIFY – Jan 2023	**Data Field % (95% CI, n)**		
	ICD-O-3	**TNM**	**Stage**
Cancer Site (n)			
Bladder and urothelial (10)	100 (69-100, 10)	70 (35-93, 7)	100 (69-100, 10)
Colorectal (141)	99 (96-100, 140)	95 (90-98, 134)	100 (97-100, 141)
Lung (83)	100 (96-100, 83)	96 (90-99, 80)	100 (96-100, 83)
Lymphoid (14)	100 (77-100, 14)	14 (2-43, 2)	100 (77-100, 14)
Oesophagus (18)	100 (81-100, 18)	100 (81-100, 18)	100 (81-100, 18)
Ovarian (13)	92 (64-100, 12)	38 (14-68, 5)	100 (75-100, 13)
Pancreas (13)	85 (55-98, 11)	77 (46-95, 10)	100 (75-100, 13)
Prostate (11)	100 (72-100, 11)	100 (72-100, 11)	100 (72-100, 11)
Uterus (30)	100 (88-100, 30)	20 (8-39, 6)	100 (88-100, 30)
Referral Pathway (n)			
Lung 2WW (91)	100 (96-100, 91)	90 (82-95, 82)	100 (96-100, 91)
Gynae 2WW (55)	98 (90-100, 54)	44 (30-58, 24)	100 (94-100, 55)
Upper GI 2WW (46)	98 (88-100, 45)	80 (66-91, 37)	100 (92-100, 46)
Lower GI clinic (150)	99 (95-100, 148)	91 (86-95, 137)	100 (98-100, 150)
Rapid diagnostic centre (38)	100 (91-100, 38)	61 (43-76, 23)	100 (91-100, 38)

b) RCRD

RCRD – Sep 2023	Data Field % (95% CI, n)	
	ICD-O-3	Stage
Cancer Site (n)		
Colorectal (106)	100 (97-100, 106)	91 (83-95, 96)
Lung (27)	100 (87-100, 27)	96 (81-100, 26)
Lymphoid (15)	100 (78-100, 15)	20 (4-48, 3)
Oesophagus (13)	100 (75-100, 13)	77 (46-95, 10)
Uterus (12)	100 (74-100, 12)	42 (15-72, 5)
Referral Pathway (n)		
Lung 2WW (27)	100 (87-100, 27)	89 (71-98, 24)
Gynae 2WW (27)	100 (87-100, 27)	52 (32-71, 14)
Upper GI 2WW (29)	100 (88-100, 29)	55 (36-74, 16)
Lower GI clinic (119)	100 (97-100, 119)	86 (78-91, 102)
Rapid diagnostic centre (24)	100 (86-100, 24)	33 (16-55, 8)

c) NCRD

NCRD – Sep 2023	Data Field % (95% CI, n)		
	ICD-O-3	TNM	Stage
Cancer Site (n)			
Bladder and urothelial (12)	100 (74-100, 12)	83 (52-98, 10)	100 (74-100, 12)
Colorectal (122)	100 (97-100, 122)	93 (87-97, 114)	100 (97-100, 122)
Lung (39)	100 (91-100, 39)	95 (83-99, 37)	100 (91-100, 39)
Lymphoid (17)	100 (80-100, 17)	0 (0-20, 0)	100 (80-100, 17)
Oesophagus (19)	100 (82-100, 19)	95 (74-100, 18)	100 (82-100, 19)
Ovarian (11)	100 (72-100, 11)	27 (6-61, 3)	100 (72-100, 11)
Pancreas (13)	100 (75-100, 13)	62 (32-86, 8)	100 (75-100, 13)
Uterus (12)	100 (74-100, 12)	25 (5-57, 3)	100 (74-100, 12)
Referral Pathway (n)			
Lung 2WW (34)	100 (90-100, 34)	94 (80-99, 32)	100 (90-100, 34)
Gynae 2WW (32)	100 (89-100, 32)	38 (21-56, 12)	100 (89-100, 32)
Upper GI 2WW (45)	100 (92-100, 45)	80 (65-90, 36)	100 (92-100, 45)
Lower GI clinic (146)	100 (98-100, 146)	86 (80-91, 126)	100 (98-100, 146)
Rapid diagnostic centre (34)	100 (90-100, 34)	47 (30-65, 16)	97 (85-100, 33)

d) DHCW

DHCW – May 2023	Data Field % (95% CI, n)	
	ICD-O-3	Stage
Cancer Site (n)		
Colorectal (20)	90 (68-99, 18)	95 (75-100, 19)
Lung (44)	87 (75-96, 39)	4 (1-15, 2)
Uterus (17)	89 (71-100, 16)	56 (33-82, 10)
Referral Pathway (n)		
Lung 2WW (52)	88 (77-96, 46)	12 (4-23, 6)
Gynae 2WW (23)	91 (72-99, 21)	57 (34-77, 13)
Upper GI 2WW (14)	79 (49-95, 11)	64 (35-87, 9)
Lower GI clinic (26)	77 (56-91, 20)	85 (65-96, 22)
Rapid diagnostic centre (7)	57 (18-90, 4)	43 (10-82, 3)

e) WCISU

WICSU – Jul 2023	Data Field % (95% CI, n)		
	ICD-O-3	TNM	Stage
Cancer Site (n)			
Colorectal (20)	100 (83-100, 20)	95 (75-100, 19)	100 (83-100, 20)
Lung (44)	100 (92-100, 44)	93 (81-99, 41)	100 (92-100, 44)
Uterus (17)	100 (80-100, 17)	24 (7-50, 4)	100 (80-100, 17)
Referral Pathway (n)			
Lung 2WW (49)	100 (93-100, 49)	90 (78-97, 44)	100 (93-100, 49)
Gynae 2WW (21)	100 (84-100, 21)	29 (11-52, 6)	100 (84-100, 21)
Upper GI 2WW (12)	100 (74-100, 12)	92 (62-100, 11)	100 (74-100, 12)
Lower GI clinic (24)	100 (86-100, 24)	88 (68-97, 21)	100 (86-100, 24)
Rapid diagnostic centre (6)	100 (54-100, 6)	67 (22-96, 4)	100 (54-100, 6)

3.5. Discordant Cases

The most common cause of discordant ICD-10 cases was an identification of anal cancer in WCISU and a corresponding identification of colorectal cancer in SYMPLIFY, which constituted 73% (n=8) of the discordant cases between WCISU and SYMPLIFY in the last data cut available (Table 13). There were no other patterns in discordant cases among the other datasets.

The most common cause of discordant ICD-O-3 morphology groupings between SYMPLIFY and the corresponding registry was a listing of cystic, mucinous, and serous neoplasms in NCRD and a corresponding listing of adenomas and adenocarcinomas in SYMPLIFY (31%, n=12) (Table 14). Among discordant ICD-O-3 morphology groupings, a listing of adenomas and adenocarcinomas was most commonly miss-paired with a listing of epithelial neoplasms in SYMPLIFY for both DHCW (60%, n=9) and WCISU (40%, n=8) (Table 14).

Analysis of discordant staging between SYMPLIFY and national registries were stratified by four groupings: the stage listed in SYMPLIFY was greater than that in the registry, the stage listed in the registry was greater than that in SYMPLIFY, the stage listed in SYMPLIFY was "uncertain" but known in the registry, or the stage listed in the registry was "uncertain" but known in SYMPLIFY. Among discordant cases, the stage reported in SYMPLIFY was often greater than that in the registries (Figure 8, Table 15). There were no instances of the national registry reporting a stage listing of "uncertain" in RCRD or DHCW, but there were in NCRD and WCISU datasets.

TNM stage concordance was low between SYMPLIFY and NCRD (51%, 95% CI = 44%-59%, n=90). However, concordance was higher for individual, T, N, and M categories at

74% (67%-80%, 130), 78 (71%-84%, 138), and 91% (86%-95%, 160), respectively, by the end of the study period (Figure 9, Table 16). Similarly, WCISU demonstrated improved concordance for individual T (65%, 95% CI = 53%-76%, n=49), N (75%, 63%-84%, 56), and M (83%,72%-90%, 62) categories compared to the overall TNM stage (49%, 38%-61%, 37). Furthermore, while TNM concordance was low, overall stage concordance was higher at 73% (67%-78%, 171) and 83% (74%-90%, 81) for NCRD and WCISU, respectively, by the end of the study period.

Table 13. Discordant ICD-10 cases based on corresponding cancer site groupings between SYMPLIFY and (a) RCRD, (b) NCRD, (c) DHCW, and (d) WCISU datasets at the final time point available for each dataset. Displayed as the number of total discordant ICD-10 cases between the registry and SYMPLIFY datasets at the last time point available for comparison.

a) RCRD (n=9)

RCRD Cancer Site – Sep 2023	SYMPLIFY Cancer Site – Jan 2023	Frequency % (n)
Oesophagus	Stomach	22 (2)
Uterus	Ovarian	11 (1)
Lymphoid	Colorectal	11 (1)
Colorectal	Anus	11 (1)
Pancreas	Colorectal	11 (1)
Colorectal	Stomach	11 (1)
Liver, bile duct	Lung, trachea, and bronchus	11 (1)
Ovarian	Uterus	11 (1)

b) NCRD (n=10)

NCRD Cancer Site – Sep 2023	SYMPLIFY Cancer Site - Jan 2023	Frequency % (n)
Oesophagus	Stomach	30 (3)
Colorectal	Anus	10 (1)
Colorectal	Prostate	10 (1)
Liver, bile duct	Unknown primary	10 (1)
Lymphoid	Colorectal	10 (1)
Lymphoid	Unknown code	10 (1)
Other	Lymphoid	10 (1)
Uterus	Ovarian	10 (1)

c) DHCW (n=4)

DHCW Cancer Site – May 2023	SYMPLIFY Cancer Site - Jan 2023	Frequency % (n)
Unknown primary	Bone and soft tissue (males), Ovarian (females)	50 (2)
Melanoma of skin	Colorectal	25 (1)
Uterus	Ovarian	15 (1)

e) WCISU (n=11)

WCISU Cancer Site - Jul 2023	SYMPLIFY Cancer Site - Jan 2023	Frequency % (n)
Anus	Colorectal	73 (8)
Stomach	Lymphoid	9 (1)
Stomach	Oesophagus	9 (1)
Unknown primary	Lymphoid	9 (1)

Table 14. Discordant ICD-O-3 cases based on corresponding morphology groupings between SYMPLIFY and (a) RCRD, (b) NCRD, (c) DHCW, and (d) WCISU datasets at the final time point available for comparison for each dataset. Displayed as the number of total discordant ICD-O-3 broad morphology grouping cases between the registry and SYMPLIFY datasets at the last time point available for comparison.

a) RCRD (n=30)

RCRD Morphology – Sep 2023	SYMPLIFY Morphology – Jan 2023	Frequency % (n)
Epithelial neoplasms, NOS	Adenomas and adenocarcinomas	17 (5)
Adenomas and adenocarcinomas	Epithelial neoplasms, NOS	13 (4)
Adenomas and adenocarcinomas	Cystic, mucinous, and serous neoplasms	10 (3)
Epithelial neoplasms, NOS	Ductal and lobular neoplasms	10 (3)
Cystic, mucinous, and serous neoplasms	Adenomas and adenocarcinomas	7 (2)
Myomatous neoplasms	Complex mixed and stromal neoplasms	7 (2)
Squamous cell neoplasms	Epithelial neoplasms, NOS	7 (2)
Blood vessel tumours	Ductal and lobular neoplasms	3 (1)
Blood vessel tumours	Epithelial neoplasms, NOS	3 (1)
Complex epithelial neoplasms	Epithelial neoplasms, NOS	3 (1)
Epithelial neoplasms, NOS	Cystic, mucinous, and serous neoplasms	3 (1)
Epithelial neoplasms, NOS	Squamous cell neoplasms	3 (1)
Lymphoid leukaemias	Adenomas and adenocarcinomas	3 (1)
Malignant lymphomas, NOS or diffuse	Mature B-cell lymphomas	3 (1)
Neoplasms, NOS	Adenomas and adenocarcinomas	3 (1)
Neoplasms, NOS	Epithelial neoplasms, NOS	3 (1)

b) NCRD (n=39)

NCRD Morphology – Sep 2023	SYMPLIFY Morphology - Jan 2023	Frequency % (n)
Cystic, mucinous, and serous neoplasms	Adenomas and adenocarcinomas	31 (12)
Adenomas and adenocarcinomas	Epithelial neoplasms, NOS	18 (7)
Acinar cell neoplasms	Adenomas and adenocarcinomas	8 (3)
Adenomas and adenocarcinomas	Cystic, mucinous, and serous neoplasms	8 (3)
Neoplasms, NOS	Adenomas and adenocarcinomas	8 (3)
Epithelial neoplasms, NOS	Adenomas and adenocarcinomas	5 (2)
Squamous cell neoplasms	Epithelial neoplasms, NOS	5 (2)
Acinar cell neoplasms	Squamous cell neoplasms	3 (1)
Complex epithelial neoplasms	Epithelial neoplasms, NOS	3 (1)
Complex epithelial neoplasms	Squamous cell neoplasms	3 (1)
Immunoproliferative diseases	Mature B-cell lymphomas	3 (1)
Lymphoid leukaemias	Adenomas and adenocarcinomas	3 (1)
Myomatous neoplasms	Complex epithelial neoplasms	3 (1)
Neoplasms, NOS	Epithelial neoplasms, NOS	3 (1)

c) DHCW (n=15)

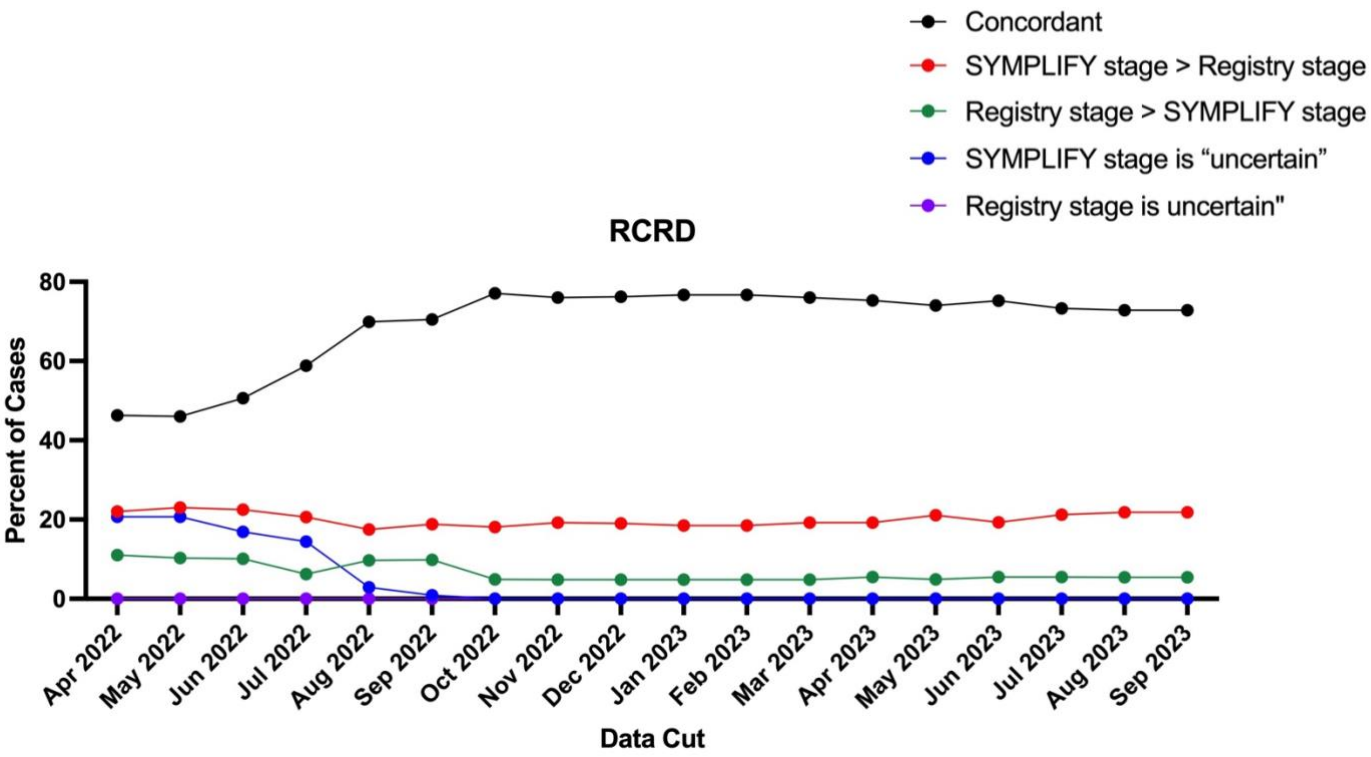
DHCW Morphology - May 2023	SYMPLIFY Morphology - Jan 2023	Frequency % (n)
Adenomas and adenocarcinomas	Epithelial neoplasms, NOS	60 (9)
Squamous cell neoplasms	Epithelial neoplasms, NOS	20 (3)
Epithelial neoplasms, NOS	Squamous cell neoplasms	13 (2)
Cystic, mucinous, and serous neoplasms	Adenomas and adenocarcinomas	7 (1)

d) WCISU (n=20)

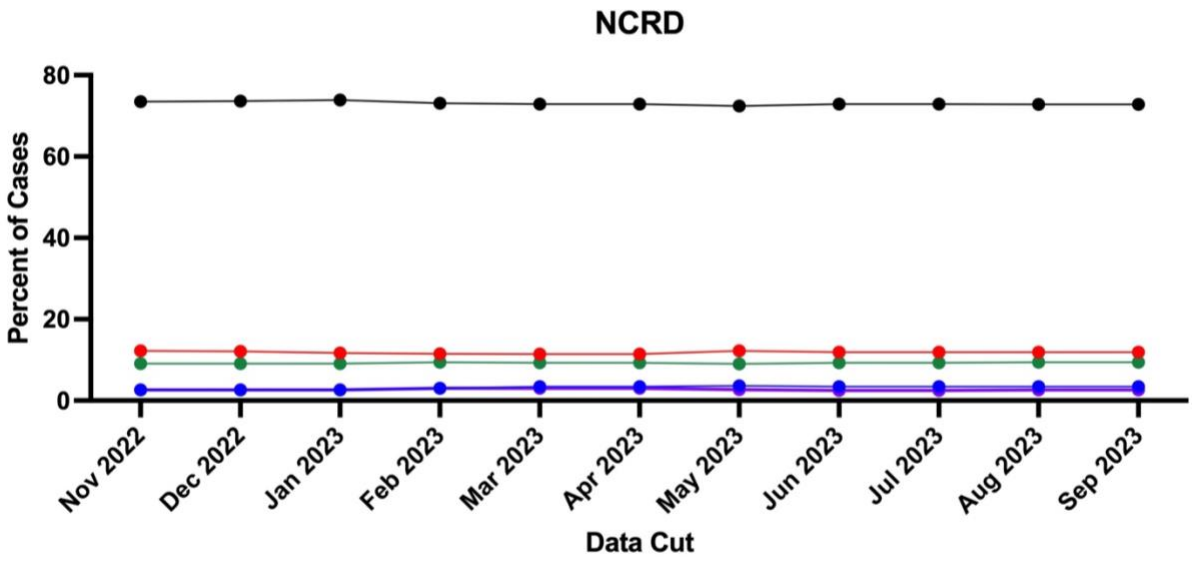
WCISU Morphology - Jul 2023	SYMPLIFY Morphology - Jan 2023	Frequency % (n)
Adenomas and adenocarcinomas	Epithelial neoplasms, NOS	40 (8)
Squamous cell neoplasms	Epithelial neoplasms, NOS	15 (3)
Acinar cell neoplasms	Epithelial neoplasms, NOS	10 (2)
Cystic, mucinous, and serous neoplasms	Adenomas and adenocarcinomas	10 (2)
Neoplasms, NOS	Epithelial neoplasms, NOS	10 (2)
Acinar cell neoplasms	Adenomas and adenocarcinomas	5 (1)
Neoplasms, NOS	Adenomas and adenocarcinomas	5 (1)
Acinar cell neoplasms	Squamous cell neoplasms	5 (1)

Figure 8. Summary of staging concordance and discordance between SYMPLIFY and (a) RCRD, (b) NCRD, and (c) DHCW and WCISU, over the study period. Displayed as a percentage of the total number of cancers considered for stage concordance in each month.

a)



b)



c)

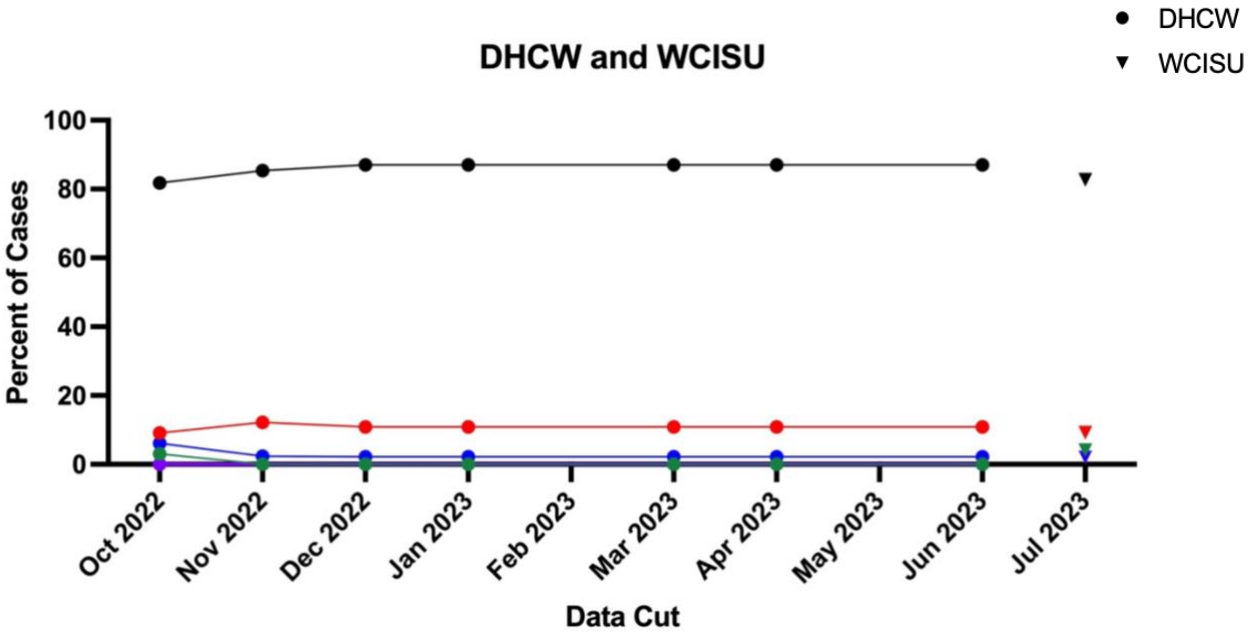

DHCW
WCISU
DHCW and WCISU
Percent of Cases
100
80
60
40
20
0
Oct 2022
Nov 2022
Dec 2022
Jan 2023
Feb 2023
Mar 2023
Apr 2023
May 2023
Jun 2023
Jul 2023
Data Cut

Table 15. Summary of staging concordance and discordance between SYMPLIFY and (a) RCRD, (b) NCRD, and (c) DHCW and WCISU, over the study period. Displayed as the number and percentage of the total number of cancers considered for stage concordance in each month.

a) RCRD

Data Cut	Total Cancers Considered	Concordant % (95% CI, n)	Discordant: SYMPLIFY stage > registry stage % (95% CI, n)	Discordant: Registry stage > SYMPLIFY stage % (95% CI, n)	Discordant: SYMPLIFY stage is "uncertain" % (95% CI, n)	Discordant: Registry stage is "uncertain" % (95% CI, n)
Apr 2022	82	46 (35-58, 38)	22 (14-32, 18)	11 (5-20, 9)	21 (13-31, 17)	0 (0-4, 0)
May 2022	87	46 (35-57, 40)	23 (15-33, 20)	10 (5-19, 9)	21 (13-31, 18)	0 (0-4, 0)
Jun 2022	89	51 (40-61, 45)	22 (14-33, 20)	10 (5-18, 9)	17 (10-26, 15)	0 (0-4, 0)
Jul 2022	97	59 (48-69, 57)	21 (13-30, 20)	6 (2-13, 6)	14 (8-23, 14)	0 (0-4, 0)
Aug 2022	103	70 (60-79, 72)	17 (11-26, 18)	10 (5-17, 10)	3 (1-8, 3)	0 (0-4, 0)
Sep 2022	112	71 (61-79, 79)	19 (12-27, 21)	10 (5-17, 11)	1 (0-5, 1)	0 (0-3, 0)
Oct 2022	144	77 (69-84, 111)	18 (12-25, 26)	4.9 (2-10, 7)	0 (0-3, 0)	0 (0-3, 0)
Nov 2022	146	76 (68-83, 111)	19 (13-27, 28)	5 (2-10, 7)	0 (0-2, 0)	0 (0-2, 0)
Dec 2022	147	76 (68-83, 112)	19 (13-26, 28)	5 (2-10, 7)	0 (0-2, 0)	0 (0-2, 0)
Jan 2023	146	77 (69-83, 112)	18 (13-26, 27)	5 (2-10, 7)	0 (0-2, 0)	0 (0-2, 0)
Feb 2023	146	77 (69-83, 112)	18 (13-26, 27)	5 (2-10, 7)	0 (0-2, 0)	0 (0-2, 0)
Mar 2023	146	76 (68-83, 111)	19 (13-27, 28)	5 (0-10, 7)	0 (0-2, 0)	0 (0-2, 0)
Apr 2023	146	75 (68-82, 110)	19 (13-27, 28)	5 (2-11, 8)	0 (0-2, 0)	0 (0-2, 0)
May 2023	123	74 (65-81, 91)	21 (14-29, 26)	5 (2-10, 6)	0 (0-2, 0)	0 (0-2, 0)
Jun 2023	145	75 (67-82, 109)	19 (13-27, 28)	6 (2-11, 8)	0 (0-3, 0)	0 (0-3, 0)
Jul 2023	146	73 (65-80, 107)	21 (15-29, 31)	5 (2-11, 8)	0 (0-2, 0)	0 (0-2, 0)
Aug 2023	147	73 (65-80, 107)	22 (15-29, 32)	5 (2-10, 8)	0 (0-2, 0)	0 (0-2, 0)
Sep 2023	147	73 (65-80, 107)	22 (15-29, 32)	5 (2-10, 8)	0 (0-2, 0)	0 (0-2, 0)

b) NCRD

Data Cut	Total Cancers Considered	Concordant % (95% CI, n)	Discordant: SYMPLIFY stage > registry stage % (95% CI, n)	Discordant: Registry stage > SYMPLIFY stage % (95% CI, n)	Discordant: SYMPLIFY stage is "uncertain" % (95% CI, n)	Discordant: Registry stage is "uncertain" % (95% CI, n)
Nov 2022	230	73 (67-79, 169)	12 (8-17, 28)	9 (6-14, 21)	3 (1-6, 6)	3 (1-6, 6)
Dec 2022	231	74 (67-79, 170)	12 (9-17, 28)	9 (6-14, 21)	3 (1-6, 6)	3 (1-6, 6)
Jan 2023	230	74 (68-79, 170)	12 (8-17, 27)	9 (6-14, 21)	3 (1-6, 6)	3 (1-6, 6)
Feb 2023	234	73 (67-79, 171)	12 (8-16, 27)	9 (6-14, 22)	3 (1-6, 7)	3 (1-6, 7)
Mar 2023	236	73 (67-78, 172)	11 (8-16, 27)	9 (6-14, 22)	3 (1-7, 8)	3 (1-6, 7)
Apr 2023	236	73 (67-78, 172)	11 (8-16, 27)	9 (6-14, 22)	3 (1-7, 8)	3 (1-6, 7)
May 2023	211	71 (64-77, 150)	13 (9-18, 27)	9 (6-14, 20)	4 (2-7, 8)	3 (1-6, 6)
Jun 2023	236	73 (67-78, 172)	12 (8-17, 28)	9 (6-14, 22)	3 (1-7, 8)	3 (1-5, 6)
Jul 2023	236	73 (67-78, 172)	12 (8-17, 28)	9 (6-14, 22)	3 (1-7, 8)	3 (1-5, 6)
Aug 2023	235	73 (67-78, 171)	12 (8-17, 28)	9 (6-14, 22)	3 (1-7, 8)	3 (1-5, 6)
Sep 2023	235	73 (67-78, 171)	12 (8-17, 28)	9 (6-14, 22)	3 (1-7, 8)	3 (1-5, 6)

c) DHCW and WCISU

Data Cut	Total Cancers Considered	Concordant % (95% CI, n)	Discordant: SYMPLIFY stage > registry stage % (95% CI, n)	Discordant: Registry stage > SYMPLIFY stage % (95% CI, n)	Discordant: SYMPLIFY stage is "uncertain" % (95% CI, n)	Discordant: Registry stage is "uncertain" % (95% CI, n)
Sep 2022	33	82 (65-93, 27)	9 (2-24, 3)	3 (0-16, 1)	6 (1-20, 2)	0 (0-10, 0)
Oct 2022	41	85 (71-94, 35)	12 (4-26, 5)	0 (0-8, 0)	2 (0-13, 1)	0 (0-8, 0)
Nov 2022	46	87 (74-95, 40)	11 (4-24, 5)	0 (0-8, 0)	2 (0-12, 1)	2 (0-12, 1)
Dec 2022	46	87 (74-95, 40)	11 (4-24, 5)	0 (0-8, 0)	2 (0-12, 1)	2 (0-12, 1)
Jan 2023	-	-	-	-	-	-
Feb 2023	46	87 (74-95, 40)	11 (4-24, 5)	0 (0-8, 0)	2 (0-12, 1)	2 (0-12, 1)
Mar 2023	46	87 (74-95, 40)	11 (4-24, 5)	0 (0-8, 0)	2 (0-12, 1)	2 (0-12, 1)
Apr 2023	-	-	-	-	-	-
May 2023	46	87 (74-95, 40)	11 (4-24, 5)	0 (0-8, 0)	2 (0-12, 1)	2 (0-12, 1)
Jun 2023	-	-	-	-	-	-
Jul 2023	98	83 (74-90, 81)	9.2 (9)	4.1 (4)	2.0 (2)	2.0 (2)

Figure 9. Concordance (%) over time between NCRD and SYMPLIFY datasets for individual T, N, and M stages, combined TNM stage, and overall stage. Displayed as the percentage of the total number of cancers considered for TNM concordance in each month.

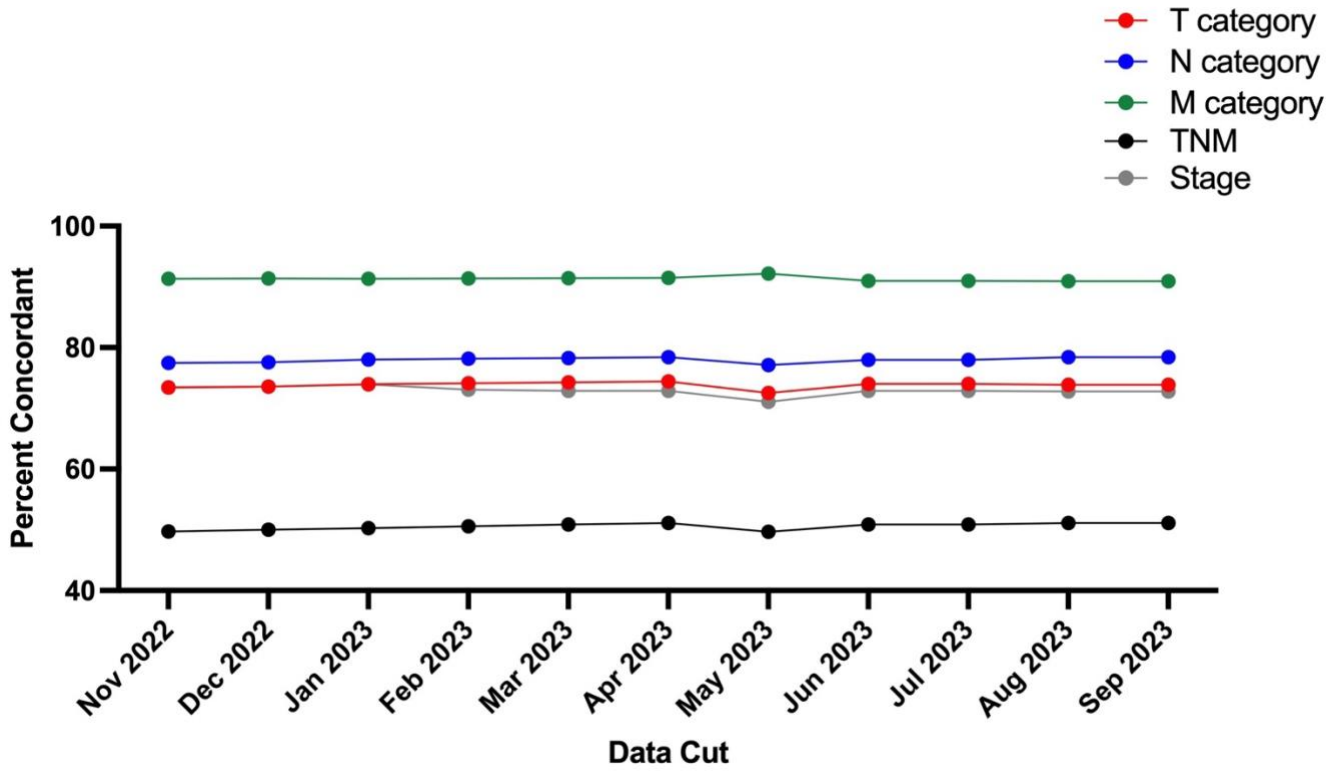


Table 16. Concordance (%) of T category, N category, M category, combined TNM stage, and overall stage between SYMPLIFY and (a) NCRD and (b) WCISU at each time point.
Proportions are based on the total number of cancers present in both SYMPLIFY and the central dataset that had TNM staging complete in both datasets at each data cut.

a) NCRD

Data Cut	T category % (95% CI, n)	N category % (95% CI, n)	M category % (95% CI, n)	TNM stage % (95% CI, n)	Stage % (95% CI, n)
Nov 2022	73 (66-80, 127)	77 (70-83, 134)	91 (86-95, 158)	50 (42-57, 86)	73 (67-79, 169)
Dec 2022	74 (66-80, 128)	78 (71-84, 135)	91 (86-95, 159)	50 (42-58, 87)	74 (67-79, 170)
Jan 2023	74 (67-80, 128)	78 (71-84, 135)	91 (86-95, 158)	50 (43-58, 87)	74 (68-79, 170)
Feb 2023	74 (67-80, 129)	78 (71-84, 136)	91 (86-95, 159)	51 (43-58, 88)	73 (67-79, 171)
Mar 2023	74 (67-81, 130)	78 (71-84, 137)	91 (86-95, 160)	51 (43-58, 89)	73 (67-78, 172)
Apr 2023	74 (67-81, 131)	78 (72-84, 138)	91 (86-95, 161)	51 (44-59, 90)	73 (67-78, 172)
May 2023	73 (65-79, 111)	77 (70-84, 118)	92 (87-96, 141)	50 (41-58, 76)	71 (64-77, 150)
Jun 2023	74 (67-80, 131)	78 (71-84, 138)	91 (86-95, 161)	51 (43-58, 90)	73 (67-78, 172)
Jul 2023	74 (67-80, 131)	78 (71-84, 138)	91 (86-95, 161)	51 (43-58, 90)	73 (67-78, 172)
Aug 2023	74 (67-80, 130)	78 (71-84, 138)	91 (86-95, 160)	51 (44-59, 90)	73 (67-78, 171)
Sep 2023	74 (67-80, 130)	78 (71-84, 138)	91 (86-95, 160)	51 (44-59, 90)	73 (67-78, 171)

b) WCISU

Data Cut	T category % (95% CI, n)	N category % (95% CI, n)	M category % (95% CI, n)	TNM stage % (95% CI, n)	Stage % (95% CI, n)
Jul 2023	65 (53-76, 49)	75 (63-84, 56)	83 (72-90, 62)	49 (38-61, 37)	83 (74-90, 81)

4. Discussion

4.1. Summary of Findings

In this evaluation of the completeness, concordance, and timeliness of cancer data collected on-site and from central databases, I highlight comparable completeness for all data fields between on-site data collected from the SYMPLIFY study and cancer data collected centrally from the gold-standard national cancer registries of England and Wales. I observed high concordance between datasets for ICD-10 and ICD-O-3 morphology groupings but lower concordance between the local and central data for stage, TNM stage, and precise morphology. However, improvements in concordance between datasets were observed for ICD-O-3 and TNM staging when broader groupings were used or when individual stages were considered. I also investigated completeness based on cancer site and diagnostic referral pathway, highlighting areas for potential improvement in data collection. Finally, I highlighted comparable timeliness between on-site and central databases. A commonly cited drawback of registry-based research is the timeliness of data entry and the completeness of data fields (8, 9). However, I show that data from the central registries of England and Wales exhibit similar completeness and timeliness to study data. This finding is promising, given the high resource burden of cancer research and the need for cost-efficient and timely data in cancer studies. To my knowledge, this is the first study to comprehensively compare the completeness, concordance, and timeliness of cancer diagnostic data collected at study sites during a prospective study in England and Wales with central registry data across many different cancer sites. Overall, this novel study highlights promising findings supporting the use of cancer registry-based diagnostic data for cancer studies in England and Wales.

4.2. Comparison with Existing Literature

4.2.1. Evaluations of English and Welsh Cancer Registries

Few studies to date have evaluated the data quality of cancer registries in England and Wales. My findings, which show a high completeness of tumour site (ICD-10) and morphology (ICD-O-3) data in the English registries, are consistent with findings in some grey literature on the topic (79). A study looking at English bowel cancer registrations between 1996 and 2004 identified a completeness rate of 60% for clinical stage among cancer registrations (78). This was lower than what I saw for both the RCRD and NCRD English datasets, in which I observed a final stage completeness of 73 (95% CI = 66%-78%)% and 100% (98%-100%), respectively. For colorectal cancers specifically, I found a very high stage completeness rate for both RCRD and NCRD at 91% (83%-95%) and 100% (97%-100%), respectively, by the end of the study period. However, it is important to note that this study was conducted before the launch of NCRAS in 2013 and occurred when cancer registrations were split up regionally rather than under a single national service (78, 79). Thus, the time frame from which this study was conducted likely explains the substantial increase I have observed in the completeness of cancer staging. Still, the study found that only four regions had stage completeness of ≥74%, and the region with the highest completeness (93%) was still lower than what I observed for stage completeness among NCRD.

A 2001 study looked at the accuracy and completeness of cancer registrations in England and Wales until May 1999. The study investigated people born in Cumbria, North West England, who had been diagnosed with cancer throughout the UK between 1971 and 1989 (88). The study found that there was 38% complete agreement between the National Health Service Central Register (NHSCR) and source data for morphology codes, while 55% of cases exhibited broad

agreement (88). This is relatively similar to my findings, in which there was an increase in morphology concordance between registry and on-site data when broad groupings were used rather than an exact match in morphology codes. However, I observed a higher overall concordance with the broad groupings, whereby concordance between on-site and registry data was 83% (95% CI = 78%-88%) for NCRD and 80% (70%-87%) for WCISU. The study also evaluated the agreement between NHSCR data and source data based on cancer site groupings. The results found that 21 (3%) cancers were true malignancies but with a significantly different diagnosis compared to what was reported in NHSCR (88). I found similar concordance between the NCRD and SYMPLIFY data for cancer site, whereby concordance was 96% (92%-98%). However, the concordance I observed for Wales was lower than this, as I found that WCISU cancer site concordance was 89% (81%-94%). The study also concluded that approximately 10% of cancer registrations were missed by NHSCR (88). This finding was very similar to what I observed in the English registry, whereby NCRD had reported 91% (87%-94%) of SYMPLIFY-England cancers by the end of the study period. Like the bowel cancer study, this evaluation was conducted before the development of NCRAS in 2013. Furthermore, the NHSCR is distinct from the cancer registries evaluated in my study, which likely contributed to the differences observed. Still, this study demonstrated similar findings to those presented here regarding case ascertainment and concordance using broader morphology groupings.

The CAP study conducted a cross-sectional comparison of the completeness and concordance of prostate cancer data collected during CAP and from NCRD. CAP found a 29.9% completeness of TNM staging in NCRD for prostate cancers, compared to 67.6% in site-collected data (80). This completeness is lower than I observed for the final TNM completeness in the NCRD dataset, which was 76% (95% CI = 71%-81%). The study also found low to

moderate agreement between CAP and NCRAS for TNM staging, with a kappa coefficient of 0.41 (80). Although I did not calculate the kappa coefficient, I similarly observed moderate agreement between SYMPLIFY and NCRD for TNM staging (51%, 44%-59%). However, the CAP study looked at prostate cancer specifically. In contrast, I investigated a wide range of cancers, with only 11 and 8 prostate cancer cases present in SYMPLIFY and NCRD, respectively, at the last data cut available for each dataset. Still, when investigating the completeness by cancer site, I found higher completeness for prostate cancers for TNM stage in both SYMPLIFY (100%) and NCRD (100%) than what was reported in the CAP study. The study also found that the completeness of stage and grade data from NCRAS improved from 2010 onwards (80). The improvements in completeness that I have witnessed in my study provide further support that the quality and completeness of registry data have improved over time.

Another study investigated the stage at diagnosis of cancers diagnosed in English residents in 2012 and followed up until the end of 2013 in the cancer registration dataset. The study indicated that missing stage rates ranged from 10.2% for lung cancers to 18.4% for prostate cancers (89). This stage completeness is lower than what I observed for NCRD, whereby stage completeness was 100% (95% CI – 98%-100%), but higher than the final stage completeness observed in RCRD, which was 73% (66%-78%). As previously outlined, the improvements in completeness that I observed in my study compared to this earlier study could be attributed to a general trend of improvement in cancer registry data quality over time.

The previous studies are the only known to date to have evaluated the completeness and concordance of cancer registry data from England and Wales using study data and since the development of NCRAS in 2013. However, annual UKIACR reports provide valuable insights

into the performance of cancer registries in the United Kingdom and Ireland. The most recent 2021 report, evaluating data from 2019, found that the completeness of tumour information for the English registry was high at 97.0%, and similarly so for the Welsh registry, at 96.9% (81). This was akin to what I observed, whereby NCRD and WCISU demonstrated completeness of 100% for both ICD-10 and ICD-O-3 data. The report found that the average staging completeness was 71.8% for England and 81.1% for Wales (81). This was substantially lower than the staging completeness I observed, which was 100% for both NCRD and WCISU. The discrepancies in these findings could be due to the special curation that registry data in the SYMPLIFY study underwent, affording higher staging completeness than what was reported by UKIACR.

The UKIACR report also highlighted that 1.5% of cases in England and 1.0% in Wales had morphological verification, but a non-specific morphology code had been provided (81). Interestingly, the percentage of cases with non-specific morphology codes increased from 2018 to 2019. Although I did not look directly at non-specific morphology codes, among the discordant morphology group cases for NCRD, 4 (10%) were instances where NCRD listed a non-specific morphology code ("Neoplasm, NOS"), while the SYMPLIFY dataset listed a more specific morphology code. Similarly, 3 (15%) discordant morphology cases between WCISU and SYMPLIFY demonstrated this same phenomenon. Thus, I found a handful of instances where the registry reported a non-specific morphology code, while a more specific morphological code was reported by the on-site data, which may reflect the increase in non-specific morphology codes reported by UKIACR.

4.2.2. Cancer Site and ICD-10 Code

In this study, I observed consistently high completeness and concordance of ICD-10 codes among all registries. This mirrors findings in Finland, New Zealand, Northern Ireland, and Scotland, where agreement between cancer registry data and medical records/site-collected data was greater than 90% (46, 55, 56, 67, 68). However, some studies have reported that the accuracy of cancer site agreement between on-site and registry data varies based on the cancer type. For example, a study comparing electronic health records found that there was substantial agreement between health records and cancer registries for cancer site-specific agreement for prostate and female breast cancer (kappa >0.60) but lower agreement for cancers such as melanoma and cervical cancer (kappa <0.40) (90). This illustrates that even though cancer site demonstrates a high agreement with on-site data, there are still discrepancies that exist among cancer types. Although I did not look at concordance by cancer type specifically, the evaluation of discordant cases highlighted that 4 (44%) discordant cancer site cases among RCRD involved colorectal cancer. In comparison, 3 (30%) discordant cancer site cases among NCRD involved colorectal cancer, and another 3 (30%) highlighted a mismatch where oesophageal cancer was indicated by the registry, while stomach cancer was identified from the on-site dataset. Furthermore, I found that the predominant cause of discordant cancer site cases in WCISU was identification of anal cancer in the registry with a corresponding identification of colorectal cancer in SYMPLIFY, accounting for 8 (73%) discordant cases at the last time point for comparison. The anatomical proximity of the anus and the colon/rectum likely explains the prevalence of this discordance between WCISU and SYMPLIFY. However, this highlights that the reporting of anal and colorectal cancers in this registry may require further attention to detail to ensure accurate reporting.

It is unsurprising that cancer site exhibited the highest concordance and completeness amongst the cancer registries and datasets I examined. Cancer types vary significantly in their activities, routes of spread, responses to treatment, and prognosis. Identifying where a cancer originated from is thus a crucial step in the clinical process to correctly determine a treatment plan and communicate prognostic information to cancer patients. Furthermore, TNM classification and overall cancer stage vary based on cancer type. For example, the T category in colorectal cancer is dependent on the extent to which the tumour has invaded the bowel wall, while the T category in pancreatic cancer is dependent largely on the size of the tumour and whether the tumour has invaded nearby blood vessels (91, 92). Thus, identifying the cancer site correctly is an essential first step in identifying other important clinical factors needed to stage, treat, and monitor the cancer. Given its importance, and the reliance of other diagnostic variables on it, the high completeness and concordance of cancer site is rather intuitive.

Cancer site information is important not only for prognostic indicators and treatment decisions but also for epidemiological studies. Public Health England runs NCRAS and uses cancer registration data to understand the epidemiology of cancer within the country (93). Ensuring the accurate reporting of cancer sites in cancer registries is important in understanding the incidence and prevalence of cancers within a country, which, in turn, influences public health strategies and programs, such as screening, resource allocation, research funding, and more. Thus, the high accuracy of cancer site reporting that I have observed in this study is important, both on a patient level and on a population level, in ensuring that cancer policy and funding are directed towards the correct cancer types.

4.2.3. Morphology and ICD-O-3 Code

The present study demonstrated high completeness of ICD-O-3 codes among all datasets, ranging from 83.6% in DHCW to 100% in SYMPLIFY-Wales, RCRD, NCRD, and WCISU. While completeness was high, concordance was more modest, with agreement between SYMPLIFY and the respective registries varying from 60% in NCRD to 74% in DHCW. However, concordance was higher when broader morphology groupings were used rather than exact ICD-O-3 codes, resulting in agreement rates ranging from 80% in WCISU to 85% in RCRD. These broader morphology concordance rates are similar to those of the FCR, where agreement between the registry and patient records for colorectal cancer histopathology was moderate (k=0.72) (55). Other studies have similarly reported discordance rates between registries and on-site data for ICD-O codes or morphology data ranging from 7.0% to 15.5% (42, 43, 68, 69). The higher end of this range is consistent with the discordance levels I observed in the broad morphology groupings. However, I observed higher discordance levels for exact ICD-O-3 codes than reported in other studies.

My analysis of discordant cases of morphology groupings shows that in many instances, the on-site and central datasets differed in level of detail rather than reporting completely different morphologies. For example, there were 12 instances where NCRD reported an ICD-O-3 code corresponding to a cystic, mucinous, and serous neoplasm, while SYMPLIFY reported an ICD-O-3 code corresponding to an adenoma or adenocarcinoma. Many serous, mucinous, and cystic neoplasms fall under the broader category of adenocarcinomas. Thus, this may represent a variation in the level of detail reported from each dataset rather than completely discordant morphology types.

Another issue complicating the assignment of morphology and ICD-O-3 codes is the significant spatial heterogeneity within a tumour. Intratumoral heterogeneity refers to the fact that different cellular populations, representing different molecular and histological properties, can exist within a tumour (94). To determine the morphology of a tumour, a sample of the cancerous lesion must be microscopically assessed. Except for in cases where an entire tumour is excised through surgical resection, the cancerous tissue for pathological assessment is obtained via a biopsy. A biopsy represents only a fraction of the tumour volume and, therefore, may not be representative of the whole tumour and the intratumoral heterogeneity. Thus, histological classifications may change after further tissue is sampled following surgical resection or when additional biopsies are taken. For example, studies examining morphological discrepancies between biopsy and resected specimens in gastric carcinoma found discordance rates ranging from 10.7% to 11.9% (95, 96). The NCRD and WCISU datasets report that the morphology variable used in this study describes the cell type of the malignant disease determined prior to the start of treatment. However, SYMPLIFY sites were encouraged to use the most up-to-date data available, so it is possible that morphology was inputted following surgical resection in some cases. Thus, different biopsies and tumour sampling methods used may have impacted the concordance of cancer morphology due to tumour heterogeneity. Alternatively, multiple histologies may be reported in the pathology report, leading to differences in CRF and registry reporting and the level of detail provided by these datasets. These represent instances where differing ICD-O-3 codes reported in on-site data and the registry may be correct, given the information available at the time of reporting. Thus, the discordance rate reported may overestimate the true discordance between the on-site and respective central datasets.

4.2.4. TNM Staging

I found that TNM staging was the variable with the lowest completeness among datasets, ranging from 74% in SYMPLIFY-Wales to 83% in SYMPLIFY-England. Furthermore, TNM staging demonstrated the lowest concordance between SYMPLIFY and the respective registries, with a final concordance of 51% in NCRD and 49% in WCISU. Although the concordance was only modest for overall TNM staging, I observed higher concordances between individual T, N, and M categories, ranging from 74% concordance for T category to 91% concordance for M category in NCRD and 65% for T category to 83% for M category in WCISU. This suggests that the discordance between on-site and central databases in TNM staging may be due to minor differences in the overall stage (e.g., a stage T2N1M0 listing versus a stage T1N1M0 listing) rather than very drastic differences in staging.

The findings of my study conflict with some other national registries regarding the completeness and concordance of TNM data. Although TNM stage demonstrated the lowest completeness of all the variables I investigated, completeness remained higher than what was witnessed in Spain, whereby completeness was 48.6%, 36.5%, and 40.0% for T, N, and M categories, respectively (58). Regarding concordance, a Swedish study that looked at prostate cancers found an overall high concordance for T category, at 83% (63). Among the three TNM variables, I observed the lowest concordance for T category at only 74% for NCRD and 65% for WCISU, contrasting with these Swedish findings. However, the low T category concordance I observed was more on par with the findings of the Swedish National Breast Cancer Register, which reported low concordance for T category at 70.1% (65). Similarly, an evaluation of the Swedish NREV found that clinical N category was one of the variables with the lowest agreement between the registry and medical records, at 79.8% (62). This is very similar to the

concordance I found for N category, which was 78% for NCRD and 75% for WCISU at the final time point available. Sweden demonstrates many similarities with the UK in health policy, healthcare provision, and their population-based cancer registry (97). These similarities are, in part, why Sweden was selected as a partner for the International Cancer Benchmarking Partnership, which was developed in England to investigate international variation in cancer survival (97). This likeness may contribute to the many similarities observed in TNM staging between UK and Swedish registries.

My study reports varied TNM completeness based on cancer site and referral pathway. TNM completeness was consistently lowest among lymphoid cancers, ranging from 0% in NCRD to 14% in SYMPLIFY. The TNM classification is typically used to assess the extent and prognosis of solid cancers, while other staging systems, such as the Binet staging system, are used for haematological malignancies, such as lymphoid cancers (98, 99). Thus, the low completeness observed for lymphoid cancers makes sense, given that these cancers utilise other staging systems. Notably, the T, N, and M variables I used for NCRD and WCISU only include UICC staging systems, which means that alternative staging methods, such as the Binet system, are not captured in this variable. The TNM variables reported in the SYMPLIFY CRF were not restricted to only UICC staging systems, leading to discrepancies between SYMPLIFY and the registries. The large proportion of lymphoid cancers in the SYMPLIFY and NCRD datasets and the inclusion of these cancers in TNM measurements, may lead to an underestimation of true TNM completeness. Similarly, cancers referred to the RDC pathway demonstrated consistently lower TNM completeness, ranging from 47% in NCRD to 67% in WCISU. Patients are referred to RDCs when they present with non-specific symptoms of cancer, which are not indicative of a particular tumour site (100). Haematological malignancies, such as leukaemias and lymphomas,

often present with non-specific symptoms, such as weight loss, night sweats, and fatigue (101). Given the symptom profile of these cancers, patients with haematological malignancies would likely be referred to RDCs, thus contributing to the low TNM completion among cancers referred to this pathway for the reasons described above.

In addition to low TNM completeness for lymphoid cancers, I also observed low completeness for ovarian and uterine malignancies. TNM completeness of uterine cancers ranged from 20% in SYMPLIFY to 25% in NCRD, while ovarian cancer TNM completeness was similarly low, ranging from 27% to 38%. Additionally, I observed low TNM completeness among cancers referred to the gynae 2WW pathway, ranging from 29% to 47% among the different datasets. The low TNM completeness among gynaecological cancers can likely be explained by the fact that there are two staging systems that can be used for gynaecological cancers. In addition to TNM staging, the International Federation of Gynaecology and Obstetrics (FIGO) has its own staging system for ovarian, endometrial, uterine, cervical, vaginal, and vulvar cancers (102). There are many similarities between the FIGO and TNM systems. However, they are distinct staging systems and are not always comparable (103). Thus, there can be confusion about which staging system to use for gynaecological cancers. A study surveyed participants in the National Gynaecological Pathology external quality assurance scheme in the UK and found that 64% of respondents used FIGO staging, 32% used both FIGO and TNM for the staging of gynaecological malignancies, and only 3% of respondents used TNM staging alone (104). Given the strong preference for FIGO staging by UK pathologists, it is unsurprisingly that TNM staging was lower among ovarian and uterine cancers in the study. Importantly, NCRD has a separate variable to input FIGO staging since only UICC data is inputted into the T, N, and M variables used in this study. Thus, in instances where FIGO staging was utilised instead of TNM, it is

likely that the TNM variable was left blank in NCRD. Future studies should investigate both TNM and FIGO staging so as not to underrepresent the staging information available due to the preference for FIGO staging among gynaecological malignancies.

Even when only the TNM classification is used, as is the case for most solid cancers, there are still challenges in assigning appropriate T, N, and M categories. The T category is determined based on the size and extent of the tumour, which involves the measurement of the tumour or the depth of invasion (105). However, difficulties in measuring the tumour and issues in the pathological interpretation of the findings, such as pesudoinvasion, remain challenges in assigning T category (106, 107). These issues likely contribute to the lower concordance I observed between on-site and central data for T category. N category refers to the lymph node involvement of the cancer, while M stage indicates whether or not there are distant metastases (105). Understanding the nodal involvement of a cancer involves radiographic interpretation. However, the diagnostic accuracy of nodal staging is dependent upon the imaging modality used and the location of the cancer (108). Nodal involvement can also be assessed via lymph node biopsies and the pathological interpretation of those tissue samples to determine if lymph nodes have cancer cells present. Thus, diagnostic accuracy and radiographic interpretation issues, as well as the availability of lymph node biopsies, may lead to differences in N category reporting. These difficulties in assigning T and N categories are important because the grading of these categories may change over time. An initial pathology report may assign a cancer as T1, or a scan may indicate no nodal involvement (N0). As new tissue samples and imaging become available, this may result in a change in the initial T and N classifications. Therefore, if TNM staging is obtained at different times between the registries and the on-site dataset, there may be discrepancies, despite the fact that the information reported by each dataset was correct at the

time of reporting. Whereas there are many different stages for T and N staging, depending on the size, extent, nodal involvement, and invasion of the tumours, M staging is more straightforward: an M category of 0 indicates no distant metastases, while an M category of 1 indicates the presence of distance metastases. Where technologies such as whole-body magnetic resonance imaging (WB-MRI) have demonstrated near 100% accuracy for skeletal and visceral metastases, there is seemingly less subjectivity and room for error in M staging compared to T and N categories (109). Therefore, it is understandable that I observed lower concordance for T and N categories but much higher concordance for the more straightforward M category.

Differences can also exist depending on whether pathological or clinical staging is used. Pathological stage utilises tissues removed during surgery to determine the stage of the cancer, while clinical stage utilises information obtained prior to surgery, such as imaging, bloodwork, biopsies, and physical exams (103). The SYMPLIFY sites were encouraged to use the best available data for staging at the time the CRF was completed. As such, it is possible that TNM staging was based on imaging alone for some cancers, while others could have used a combination of clinical and pathological data. Thus, it is unclear whether SYMPLIFY reported clinical or pathological stages. As for NCRD and WCISU, the stages listed are integrated stages using all the information available at the time the registration is finalised. Thus, these discrepancies in how the TNM data was collected from the site and registries likely contributed to the TNM concordance rates I observed.

4.2.5. Stage

I reported an overall high level of completeness of stage in SYMPLIFY and the gold standard national cancer registries, with completeness at 100% for SYMPLIFY-England,

SYMPLIFY-Wales, NCRD, and WCISU. However, I demonstrated comparatively lower stage completeness for the rapid central datasets, with a final stage completeness of 73% (95% CI = 66%-78%) for RCRD and 43% (34%-53%) for DHCW. The stage completeness I observed for the gold standard registries was higher than what was found in Finland, New Zealand, and Spain, where missing stage data ranged from 11% in the FCR to 62.1% in Spain (44, 55, 58). Furthermore, I demonstrated moderate agreement between SYMPLIFY and the respective registries for stage, ranging from 73% in NCRD and RCRD to 87% in DHCW. This concordance is comparable to findings in other cancer registries, whereby agreement between the registries and medical record data was moderate for Finland, and discordant rates ranged from 11.2% to 20% in New Zealand, the United States, Canada, and Denmark (43, 46, 48, 53, 55). Thus, the central databases evaluated in my study demonstrate staging agreement on par with other national cancer registries worldwide.

Among the databases with lower stage completeness, namely RCRD and DHCW, I observed discrepancies in stage completeness based on cancer site and referral pathway. In RCRD, I reported low stage completeness for lymphoid and uterine cancers at 20% and 42%, respectively. As such, the RDC and gynae 2WW pathways also demonstrated the lowest stage completeness at 33% and 52%, respectively. Although DHCW did not have enough lymphoid cancers to investigate completeness by that cancer site, there was similarly low stage completeness of uterine cancers at 56%. Interestingly, DHCW demonstrated very low completeness for lung cancer staging, at only 4%. These trends were mirrored in the completeness by referral pathway in DHCW, whereby lung 2WW had the lowest stage completeness at 12%, followed by RDC and gynae 2WW at 43% and 57%, respectively. However, this issue was remedied in the gold standard registry, whereby stage completeness of

lung cancers in WCISU was 100%. These differences may reflect disparities in data access and acquisition between DHCW and WCISU, whereby DHCW may not have had access to all necessary staging information, such as radiology reports, pathology reports, and multi-disciplinary team notes, which would be necessary for complete and accurate staging.

The lower reporting of stage for lymphoid cancers could be attributed to the larger issue of underreporting of haematological malignancies, including leukaemia and lymphoma, which has been reported in other studies (57, 64). Previous studies have found that haematological, CNS, and soft tissue cancers are more likely to be missed in the registry. As such, it is not far-fetched to suggest that when these cancers are, in fact, registered, they may be more likely to have missing data compared to other cancer types. The low completeness of staging for uterine cancers could be the result of poor TNM staging completeness for gynaecological cancers highlighted in the previous section.

The low stage completeness observed among lung cancers in DHCW is not what would be expected, given the findings of previous studies and the epidemiology of lung cancer. In a study looking at the stage completeness in Mallorca, Spain, overall completeness was low but was higher among some cancer sites, including lung (58). Meanwhile, the NPOCR highlighted high case completion and data accuracy for certain cancer types, including lung and bronchial malignancies (48). Furthermore, lung cancer is the third most common cancer in the UK, accounting for around 20% of cancer deaths (110). The high prevalence and burden of disease associated with lung cancer would lead one to expect that the completeness of cancer variables would be high like other high-prevalence cancers, such as colorectal and breast. However, despite the significant burden that lung cancer places on individuals and populations, certain demographic factors may contribute to the low stage completeness witnessed in DHCW. For

example, several studies have shown that there is a general underreporting of cancers, particularly stage, in older patients and those with higher comorbidity (45, 52, 64, 111). Given that lung cancer is more frequently diagnosed in older adults, with over 40% of cases in the UK being diagnosed in those 75 years and older, age and comorbidity status may be contributing to the lack of stage reporting seen in the DHCW dataset (110). Still, a previous paper comparing study data to FCR data found that pancreatic and lung cancers were among the cancer types with the longest delays to registration, as it took an average of 1.7 years for 95% of lung cancers to be registered (56). Thus, the lack of lung cancer staging information available in DHCW may, therefore, reflect longer delays in obtaining diagnostic cancer information that has been mirrored in studies of other cancer registries. Nonetheless, the low stage completeness of lung cancers observed in DHCW is an area for potential improvement moving forward.

Interestingly, I found that when there were discordant stage listings between SYMPLIFY and the respective registries, most of the time, the stage listing in SYMPLIFY was higher than that of the registry. I found that a stage listing higher in SYMPLIFY than in the registry occurred 22%, 12%, 11%, and 9% of the time for RCRD, NCRD, DHCW, and WCISU, respectively. Meanwhile, the opposite occurrence, a stage listing higher in the registry than in SYMPLIFY, occurred 5%, 9%, 0%, and 4% of the time, respectively. I had hypothesised that in the event of discordant stage cases between SYMPLIFY and the respective registries, the stage listing would be higher in the cancer registry than in the study data. This is because there is often an increase in cancer staging following the initial diagnosis due to further imaging and diagnostic tests. As such, I predicted that when the SYMPLIFY cancers were recorded at the limited 3- or 9-month cut-offs, all staging information might not have been available then. Thus, the longer and more comprehensive follow-up that registry data allows for would have had a better insight into all

staging information, which would tend to lead to a higher stage than what was reported in SYMPLIFY in the event of discordant cases. This was observed in a comparison of cancer registry data to patient charts for germ cell tumors in Canada. The study found that among the 17.3% of discordant cases, the stage that was reported in the registry was higher 11.4% of the time, compared to only 4.3% of cases where the stage was higher in the patient charts (43). However, this was the opposite of what I found.

The likelihood that a cancer stage is reduced after further investigations are conducted is relatively low. This is because further tests, such as whole-body imaging (e.g., positron emission tomography (PET) scans and full-body computed tomography (CT) scans), can uncover more extensive malignancies than once previously thought. Thus, it is much more likely that, say, a stage III cancer will be upgraded to stage IV after further investigations than being downgraded to stage II. Downstaging may occur after treatments have been initiated, but the stage at initial diagnosis reflects the stage prior to any treatments. As such, there is reason to believe that in the instances where the cancer stage listed is higher in SYMPLIFY than in the registry, the stage listed in SYMPLIFY is more accurate. Previous studies have reported that locally advanced or metastatic cancers tend to have higher proportions of missing or inaccurate stage data than localised cancers in cancer registries (44, 53, 55). These findings align with what I observed, whereby higher staged cancers, according to the on-site dataset, were more frequently reported as lower stages in the registry. A study of the NZCR found that the inaccurate and missing reporting of stage data due to demographic, treatment, and other factors overall led to an underestimation of metastatic breast cancer in the registry by around 21% (44). These studies illustrate that registries may tend to underreport advanced-stage cancers, which aligns with what I observed in my study, whereby the cancer stage tended to be higher in the on-site data when

there were discrepancies between the two datasets. Thus, my findings may also point to a tendency for underreporting of advanced-stage cancers among the UK registries, which could have significant effects on epidemiological studies. Future research should investigate the factors that may contribute to stage discrepancies and further investigate the potential underreporting of advanced-stage cancers.

Staging accuracy is important for both epidemiological surveillance and for the monitoring of cancer benchmarks. As part of NHS England's goal to improve early cancer diagnosis, they have set out a goal to diagnose 75% of all cancers at an early stage (i.e., stage I or II) by 2028 (112). Central cancer databases in England, such as NCRD and RCRD, will play a key role in evaluating this outcome and determining whether this goal is achieved. I found that both RCRD and NCRD demonstrated concordance with SYMPLIFY of 73% by the end of the study period, indicating that both demonstrate comparable suitability for stage monitoring. My study found that registries tended to report a lower cancer stage than SYMPLIFY in times of discordance. However, it is essential to better elucidate the impact of these different cancer stages. For example, the difference between one dataset reporting a stage I cancer and the other reporting a stage II cancer has fewer implications on public health monitoring than one reporting a stage I cancer and the other reporting, say, a stage III cancer. In the first example, both are still reporting early-stage cancers. In contrast, in the second example, one dataset is reporting an early-stage cancer and the other an advanced-stage cancer. Future studies should better elucidate the differences observed between on-site and registry datasets in stage reporting to determine whether there are significant differences between these datasets in the reporting of early- and advanced-stage cancers. It is important to understand this because inaccuracies in staging

information in these datasets may have broader implications for monitoring this NHS early diagnosis goal.

While the concordance between SYMPLIFY and the respective registries was moderate, it is again important to highlight the discrepancies in how stage was reported in the various datasets. As described previously, hospital sites were encouraged to use the best available evidence at the time of reporting, so it is possible that a combination of pathological, clinical, and integrated stage information was used to determine the stage in the on-site dataset. Whereas NCRD and WCISU reported an integrated stage. The stage reported in RCRD is a combination of pre-treatment or pathological stage, either as reported by the sites or calculated from the TNM components. This illustrates that the methods of staging utilised by the various datasets were not consistent and, therefore, may have impacted the discordance rates for stage. Future studies should ensure consistency between stage reporting (e.g., pathological, clinical, integrated) to allow for fair data comparisons.

4.2.6. Completeness and Timeliness of Cancer Registrations

I observed overall high completeness of cancer registrations in the national cancer registries and comparable timeliness between the national registries and the on-site SYMPLIFY dataset. I found that 91% of the SYMPLIFY-England cancers were registered within NCRD by the last data cut available, while 77% were registered in RCRD. Similarly, I found that 81% and 88% of SYMPLIFY-Wales cancers were registered in WCISU and DHCW by the end of the study period. I further investigated the 23 cancers diagnosed in SYMPLIFY-England but not reported in NCRD. Upon investigation, I found that 14 of these 23 cancers were, in fact, registered in NCRD but had been excluded from the study due to a date of diagnosis preceding

the date of enrolment or due to an ineligible ICD-10 code being reported in NCRD. After including these cancers, the registration completeness rate for NCRD improves to 97%. The European Network of Cancer registries outlines a strict hierarchy of points indicating how the date of diagnosis should be reported in cancer registries to allow for comparability between registries (113). This epidemiological definition of date of diagnosis may differ from what is reported in clinical data, where the date of diagnosis follows less rigid criteria. The differences between these methods could explain why some cancers were eligible for inclusion according to the date of diagnosis outlined in the SYMPLIFY dataset but not in the NCRD dataset. As for the remaining 9 cancers, they were indeed "missed" by NCRD. However, 6 of these cases involved patients who had historical cancers reported, and 2 involved a patient with multiple primary cancers. Cancer registries are interested in reporting new primaries and typically do not report new cancer registrations for the recurrence of previous cancers. Meanwhile, a study like SYMPLIFY is interested in all cancer diagnoses because the goal is to detect whether cancer is present at all, regardless of the nature of that cancer. Thus, in those instances where cases were "missed" by the registries but reported in SYMPLIFY, it is possible that at least some were instances of recurrent cancers, which would be reported by the study dataset but would not necessitate a new cancer registration in the various registries.

Prior studies have shown that older patients, typically over the age of 70, and patients with more comorbidities experience a relative underreporting in cancer registrations or have less complete data available (45, 52, 64). Although I could not look at patient demographic information and comorbidities, the fact that most of the cancers that NCRD missed occurred in patients with historical cancers or multiple primaries suggests that similar factors might be at play. Factors such as previous cancer diagnoses or multiple primary cancers might add a layer of

complexity to the cancer registration process, which may increase the rate of missed cancers compared to those in which a single and first cancer diagnosis is being made. However, a previous study that looked at a sample of multiple cancer registrations in England and Wales from 1971-1980 found that only 61% of cases were true occurrences of multiple primaries according to the cancer registries, while the rest were deemed to be reporting errors (e.g., duplicate registration of a single tumour) (114). Therefore, it is unclear if, in these instances, the registration was truly missed by the registry or if multiple primaries were reported in error by the sites. To try and avoid this issue, if multiple cancers for a patient with identical information for ICD-10 code, stage, date of diagnosis, and morphology were reported, these were considered duplicates and one of the entries was removed. Still, future research should investigate whether there is a lack of reporting of cancers in patients with multiple primaries or historical cancers in cancer registries in England and Wales.

My findings are similar to other reports of case ascertainment in other national cancer registries, which range from 88.0% to 99.2% (45, 48, 57, 60, 64, 67). Previous studies have found that certain cancer types, including soft tissue, CNS, and haematological malignancies, experience a relative lack of reporting in cancer registries, while other common cancer types such as lung and bronchus, colorectal, and female breast cancers demonstrate high completeness of cancer registrations (48, 57, 60, 64). Interestingly, I found that amongst all the registries, the cancer sites most commonly associated with missed cancers were either colorectal or lung, trachea, and bronchus. However, this is likely due to the recruiting strategy, whereby lower GI and lung 2WW pathways were used to recruit patients, and the high prevalence of these cancer types. Looking at the percentage of missing cancers based on the total number of cancers per site helps to reduce the skewing of results based on cancer prevalence and recruitment strategy.

When doing this, I found that less common cancer sites, such as thyroid, cancers of unknown primary, and bone and soft tissue cancers, had high proportions of missing cancers in the respective registries. Interestingly, I found that in RCRD, there was a high proportion of pancreatic cancers that were not registered (89%, n=8). However, this was remedied in the gold standard registry, whereby no pancreatic cancers were missed in NCRD. This may reflect previous findings that pancreatic cancers take longer to be registered, and therefore, were missed by the rapid dataset but were registered in the gold standard registry (56).

While I found that haematological and CNS cancers were not frequently missed in the various registries, as previous research has reported, I found that close to a third of lymphoid cancers and all plasma cell cancers that were reported in the gold standard national registries were not reported in SYMPLIFY. Melanoma of the skin was another cancer site that was relatively underreported in the on-site dataset. These findings suggest that cancer studies looking specifically at these malignancies may have better case ascertainment using centralised data.

An interesting finding of my study was that there were cancers registered within the national registries that had not been reported in SYMPLIFY, and many of these cancers were diagnosed within the mandatory follow-up period for the SYMPLIFY study. This finding suggests that cancer registry data may afford complete and more thorough follow-up of study patients than is feasible with data collected on-site. This idea is in line with what was observed in the TASTE trial, whereby the use of registry data resulted in no patients being lost to follow-up—a considerable feat for any trial (15). Thus, these findings suggest that registry data may be superior to data collected on-site in terms of case ascertainment.

I used completeness and concordance compared to the final dataset available as a measure of timeliness of each dataset. I found that SYMPLIFY reached full completeness of

cancer registrations compared to the final dataset at 12 months following the end of the study enrolment. NCRD and RCRD both reached completeness of cancer registrations with only a one-month delay at 13 months following enrolment. DHCW also showed comparable timeliness to SYMPLIFY, reaching completeness of registrations at 15 months post-enrolment. Concordance followed similar patterns, indicating that once the data is complete, it is also quite accurate compared to its final dataset. An investigation of the timeliness of the SCRCR found that 98% of cancers were enrolled in the cancer registry within 12 months of enrolment, while 95% of cancers were diagnosed within the same time frame in the NPCR for Sweden (60, 63). These timelines are similar to what I have witnessed in my study, in which the completeness of cancer registrations ranged from 12-15 months for the various datasets. However, some studies have reported timelier registrations than this. In comparing study data to registrations in the FCR, it was found that most cancers were registered within 0.9 months (56). Meanwhile, the median time to registration in the Swedish NREV was found to be 3.9 months from diagnosis, while that in the CRN was 261 days in 2005 (57, 62). The first data cut for SYMPLIFY and RCRD came at 5 months following enrolment, while that of DHCW, NCRD, and WCISU were from 10, 12, and 20 months following enrolment, respectively. Thus, I cannot decipher how many cancers were registered within the timeframes explored in these prior studies. As well, without data on the date of registration, I was unable to calculate the average time between diagnosis and cancer registration. Further, since I was using the end of the enrolment period rather than the actual date of diagnosis as the comparator for timeliness, this may have underestimated timeliness, as some patients may not have been diagnosed with cancer until weeks to months after the enrolment period ended. Despite these limitations, my study still demonstrated a high level of completeness

of cancer registrations at approximately 1 year following enrolment, which illustrates comparable timeliness to evaluations of other European cancer registries.

The timeliness of cancer registrations has increased steadily over the past two decades due to technological advancements. Previously, the cancer registration process would require manual extraction of cancer data from hospital records, leading to a longer lag time to registration and more opportunities for missed registrations. Over the years, cancer registration has become a much more efficient process, with registries using informatics to pull the necessary cancer data directly from electronic health records (115). While my study shows that cancer registry data can be obtained in a timely manner comparable to on-site data collection, with further advancements in technology and the potential future use of artificial intelligence, it is possible that cancer registration will continue to become even more efficient and timely.

It is important to think about completeness in terms of timeliness as well. It is possible that those cancers that were deemed to be "missed" in the registries were merely experiencing a longer lag time to registration. For example, the FCR reported that 95% of cancers were registered within 0.9 months of diagnosis, but there were variations in the length of time to registration (56). The study found that lung cancers experienced a longer lag time in registration at 1.7 years, while it took 3.2 years for 95% of pancreatic cancers to be registered (56). However, a wide range in the completeness of cancer registration has been reported among European cancer registries, with a median time to completion of 18 months and a range of 4-60 months (116). Factors that have been previously mentioned, such as comorbidities, age, treatments, multiple primaries, historical cancers, and more, may lead to certain cancers experiencing a much longer delay between diagnosis and enrolment. Given that my study stretched only until 22 months following the end of the enrolment period, it is very plausible, based on prior reports,

that some of the cancers reported in SYMPLIFY that were not reported in the cancer registries may, in fact, be registered in the future. Additional studies should institute a longer follow-up period for cancers that were not registered to determine if the registries truly missed them or if they are experiencing longer delays in registration.

4.3. Strengths and Limitations

This is the first study to compare diagnostic data from a prospective cohort study to cancer data in central registries in England and Wales over many cancer sites and since NCRAS was developed in 2013. The comprehensive nature of this work, which included most cancer types with only narrow exclusion criteria, was a strength, as many previous studies of cancer registry data quality have focused only on a subset of cancers. As such, the study provides an encompassing overview of cancer registration completeness, concordance, and timeliness among the datasets investigated. The recruitment of patients from central and peripheral hospitals across the country also provided a representative sample of cancer registrations, as geographic discrepancies in cancer registrations have been reported in the past.

Despite these strengths of the study, several limitations remain. The relatively small sample size, which looked at fewer than 500 cancers in each dataset, is a limitation compared to other studies of cancer registry data quality which utilise thousands of cancer registrations. The population of cancers evaluated for concordance with SYMPLIFY was even smaller, as I could only assess cancer registrations that were reported in both the on-site and central datasets. The small sample size may not be as representative of the English and Welsh cancer registries compared to larger samples. Furthermore, the small sample size limited the ability to assess the data quality of cancer registrations for rarer cancers such as kidney, brain, thyroid, and more

(117). However, the relatively small sample size allowed me to achieve greater resolution by analysing each of the cancers in much more depth than would have been possible with a much larger dataset.

The timeliness of the data cuts, whereby I received the first data cut from each dataset at 5 months following enrolment for SYMPLIFY and RCRD and 10, 12, and 20 months following enrolment for DHCW, NCRD, and WCISU, respectively, prevented the ability to assess how many cancers are registered very rapidly (i.e., within the first 3 months of enrolment) as other previous studies have assessed. The study timeline also did not allow me to evaluate cancers that have a very long delay in registration, as I only received data cuts up to 22 months following enrolment, while prior studies have noted median cancer registration times of more than 3 years following diagnosis. Timely data acquisition is invaluable in cancer research and epidemiological monitoring. Therefore, being unable to assess cancers that were registered before the first data cuts I received and those that experienced long delays in registration limited my ability to investigate the true timeliness of data acquisition. Using a longer follow-up period that starts immediately after enrolment and continues for years will allow for better elucidation of timeliness and those cancers that experience cancer registration timeliness on one of the extremes.

As previously discussed in the TNM staging and stage sections, differences in staging methodology used between the different datasets were a limitation. There were inconsistencies in how the different datasets reported stage and discrepancies in which data was available when stage was recorded. These discrepancies likely resulted in a higher discordance rate than if each dataset used the same methodology. Therefore, future studies should ensure that equivalent stage variables are being considered to allow for a fair comparison between datasets. However, the

flexible approach I used in allowing various staging methods is more reflective of inter-registry comparisons, where different national registries worldwide utilise different rules regarding stage. Despite these inconsistencies, I still found moderate stage concordance between SYMPLIFY and the central datasets.

Another limitation with regards to stage was that the TNM variables used in NCRD and WCISU did not encompass staging systems outside of UICC 8th edition, while the SYMPLIFY CRF allowed for any staging system to be inputted and identified when filling out the TNM variable. As a result, this may have led to an underestimation of TNM staging completeness in the NCRD and WCISU datasets. However, this highlights the flexible approaches that can be taken in collecting study data compared to registry data, allowing for comprehensive data acquisition with fewer variables.

An important limitation of this study is the assumption that the SYMPLIFY data are correct in times of discordance between SYMPLIFY and the various registries. This assumption was made to facilitate comparisons between the on-site and registry datasets. Still, it is not necessarily true that in times of discordance, the study data are always correct. As I have discussed extensively throughout this thesis, there are many things to consider when assessing the discordance rates reported in this study, such as differences in the time of reporting between the various datasets, the different staging methodologies used, and the level of detail reported. Furthermore, the gold-standard national registries (NCRD and WCISU) undergo a rigorous registration process whereby at least two cancer registration officers must verify the data for a given cancer registration. Thus, there is reason to believe that the registry data may be more likely to be correct in times of discordance. The assumption that the SYMPLIFY data was correct was necessary to evaluate concordance as a surrogate marker for accuracy in this study,

but this assumption is not necessarily true, and we would expect the registry datasets to be correct in at least some cases of discordance. Future studies comparing study data to registry data should take a sample of discordant cases and compare the data obtained to patient medical records to determine the percentage of discordant cases where the study data was correct versus the registry data.

An important consideration when evaluating the results of this study is that all participants were referred by their primary care physicians to one of five urgent referral pathways for rapid investigation of symptoms concerning for cancer. While it is true that cancer diagnoses via urgent referral pathways represent nearly 40% of cancer diagnoses in the UK, the findings of this study may not be applicable to cancers that are diagnosed via other routes, such as through emergency presentations or routine screening (118). Furthermore, the findings may not be comparable to other evaluations of cancer registries worldwide that investigate registry data quality among cancers diagnosed via means other than urgent referral pathways. It is important also to consider the impact of the recruitment strategy on the cancers that were diagnosed in my study. Although all cancer types were included in the present study, except non-melanoma skin cancers, patients were recruited via rapid diagnostic, lung, upper GI, lower GI, or gynaecological 2WW pathways, resulting in a higher proportion of cancer types such as lung, colorectal, ovarian, etc. Thus, my findings may be most applicable to these common cancer types given the recruitment strategy and higher proportion of certain cancer types.

Finally, a crucial limitation of my study is the differences in resourcing and engagement between the bodies overseeing the different datasets investigated. Due to resource limitations, only a single data cut was obtained from WCISU, which prevented the ability to assess timeliness in this dataset and to get a more holistic view of the changes in data quality over time.

Due to these same limitations, WCISU could only provide data on cancer registrations that had been previously identified in DHCW data. As a result, this may have impacted the case ascertainment numbers identified in WCISU. Future studies should ensure adequate data cuts from each dataset to ensure comparisons over time can be investigated. Although cancer registries have detailed information on all the cancers diagnosed within a population, curating and providing this data for research purposes is not within the standard scope of registry services. Thus, resourcing, funding, and staffing outside of the standard registration process are necessary to use registry data effectively in research. In this study, the English registries were supported earlier and to a greater extent than the Welsh registries. Furthermore, the NCRD and RCRD datasets were specifically curated for the SYMPLIFY study due to funding and resourcing support provided by the study organisers. However, similar resourcing was not provided to the Welsh registries, which likely contributed to the discrepancies in data availability between the English and Welsh datasets. Thus, although my findings demonstrate that complete and accurate registry data is achievable in a timely manner, it is crucial to note that this is dependent on the resourcing and capacity of the cancer registries. In other words, the level of data quality and timeliness presented here may not be available for every study.

4.4. Implications and Future Work

This study demonstrates promising findings of comparable completeness, concordance, and timeliness between data collected on-site during a prospective cohort study and central data collected from cancer registries in England and Wales. These results encourage the use of central registry data in cancer studies conducted in England and Wales.

The Gunnarsson paper suggested criteria to help determine whether a dataset is adequately complete and valid. These criteria included completeness of variables exceeding 95%, the number of missing registrations not exceeding 10%, and a discordance rate below 5% (41). The rapid datasets evaluated in this study, RCRD and DHCW, had missing case registration rates of 23% and 12%, respectively, which exceeds the criterion previously outlined. As for the gold standard registries, NCRD had a missing case registration rate of 9%, which was reduced to only 3% after further investigations identified that some cancers had been registered but removed from my analyses due to our exclusion criteria. In addition to having a high case completion, NCRD had high completion of variables, which was 100% for ICD-10, ICD-O-3, and stage, but only 76% for TNM stage. All variables but for TNM stage satisfy the criterion of variable completeness exceeding 95%. As for discordance, there was very high concordance between NCRD and SYMPLIFY for ICD-10 at 96% but lower for ICD-O-3 morphology code, broad morphology, TNM stage, and overall stage. However, as I have described extensively, the discordance rates cannot be taken at face value. Factors such as the utilisation of different staging systems, differences in the level of detail reported, and variation in the time of reporting likely resulted in an inflated discordance rate between the SYMPLIFY and registry datasets. Given these findings, it is clear that the NCRD dataset satisfies two of the criteria outlined in the Gunnarsson paper. Additionally, ICD-10 data satisfied the <5% discordance criterion, and there is reason to believe that the true concordance between the SYMPLIFY and NCRD data is higher than what I have reported. Thus, based on these outlined criteria, I think there is sufficient evidence to say that NCRD data is complete and valid compared to pre-specified cut-offs.

WCISU had a similarly high completion of variables, which was 100% for ICD-10, ICD-O-3, and stage, but only 77% for TNM stage. Like NCRD, several factors impacted the

concordance measures between WCISU and SYMPLIFY, making it hard to comment on whether the criterion of <5% discordance was achieved. Still, WCISU had a missing registration rate of 19%, which exceeds Gunnarsson's cut-off of <10% required for complete and accurate data. Unfortunately, I was unable to investigate the causes of these missed cancer registrations as I did for NCRD, so it is possible that the missing cancer registration rate is less than what I have reported here. These findings suggest that NCRD likely meets the criteria for complete and accurate data outlined in the Gunnarsson paper, but further investigation into case ascertainment in WCISU is necessary to determine if this is true for the Welsh dataset. The RCRD and DHCW datasets did not meet these cut-offs, which further supports the use of rapid datasets as a quick snapshot of the cancer registration process rather than a source of complete and accurate data to be used in research.

Given these findings regarding the validity of the registry data, it is important to consider how the results of the SYMPLIFY study would have changed if registry data had been used rather than the on-site dataset. As described above, the NCRD dataset had a high case completion of approximately 91%. However, an additional 55 cancers were reported by NCRD that were not reported in SYMPLIFY, resulting in 291 cancers reported in NCRD compared to only 259 in SYMPLIFY. Using the NCRD dataset would have resulted in more cancer diagnoses reported throughout the study, which would have impacted the reported positive and negative predictive value of the MCED test evaluated. As for RCRD, only 226 cancers were reported by the registry, compared to the 259 English cancers. Using this rapid registry would have resulted in missing cancer diagnoses. This finding is similar to WCISU, where only 112 Welsh cancers were reported, compared to 121 in SYMPLIFY, while DHCW reported more cancers in SYMPLIFY at 122.

In addition to the number of cancers diagnosed, it is important to consider the impact of the concordance between SYMPLIFY and the cancer registries on the original study's results. The SYMPLIFY study was not only concerned with whether a cancer was present to determine the ability of the MCED test to detect cancer, but it was also concerned with the cancer site. The MCED test in question reports a cancer signal origin, which is the most likely site where the suspected cancer arose. As such, the ICD-10 code plays a vital role in determining the accuracy of this MCED test function. NCRD and RCRD demonstrated a high concordance with SYMPLIFY for ICD-10 at 96% and 95%, respectively. Given the high concordance, it is likely that the use of these registry datasets would not have resulted in substantially different results regarding the accuracy of the MCED cancer signal origin prediction. As for the Welsh registries, ICD-10 concordance with SYMPLIFY was 96% and 89%, respectively, for DHCW and WCISU. Again, the high concordance with SYMPLIFY for DHCW likely would not had a significant effect on these findings, but at 89%, WCISU may have had a more significant impact on the study results. Another critical data point in the original SYMPLIFY study was stage, as the authors determined that the MCED test sensitivity increased with stage. With concordance rates ranging from 73% in RCRD and NCRD to 87% in DHCW, the lower concordance between the various registries and SYMPLIFY for stage would have likely impacted these findings. Furthermore, the lower stage completion observed in RCRD (73%) and DHCW (44%) would have likely limited the ability to assess the impact of stage on MCED sensitivity at all.

To truly understand the impact of using cancer registry data for trial follow-up future studies should investigate precisely how the use of the different datasets would have impacted the results of the SYMPLIFY study and consider how study factors (e.g., study size, length of follow-up, etc.) would need to be altered if cancer registry data were used to prevent significant

differences in study results compared to using on-site data. Providing guidance on how to design a study to ensure that the use of cancer registry data can be used in clinical trial design without producing significantly different results from those produced from on-site data collection is necessary to allow for the widespread adoption of registry-based research.

Despite these unknowns about the impact on study findings, registry-based research is becoming increasingly attractive due to the availability of cancer registry data and the resource burden associated with cancer research. Registry-based trials, where registries are used in the recruitment, data collection, and/or follow-up of patients, are one way that registries can be used to support research (8). Pragmatic trials and real-world evidence studies are also growing in popularity due to their reduced burden on participants, real-world applications, and improved generalisability, which can also benefit from the use of cancer registry data (119-121). Several studies have highlighted the potential benefits of using cancer registry data for research, including improved generalisability, ease of recruitment, reduced burden on patients due to less intrusive follow-up, and, most notably, decreased costs (8, 11-17, 20). Much money is spent globally on cancer research, and the resource burden associated with cancer is particularly prevalent in the UK. The UK Health Research Analysis 2022 reported that £5 billion is spent on healthcare research, and nearly a fifth of this is dedicated to cancer research (122). However, the proportion spent on cancer research over time has declined from 20.4% in 2004 to 16.8% in 2022 (122). Despite this funding, with the rising number of cancer cases, Cancer Research UK estimates there will be a funding gap of over £1 billion for cancer research over the next decade (123). These findings highlight the urgent need to make cancer research more resource-efficient. Improving the cost and efficiency of cancer research will not only allow for further

advancements in highly prevalent cancers but will also hopefully allow for the diversion of resources to rarer and less well-studied malignancies.

Given the high resource burden associated with cancer research and the gap in cancer research funding in the UK, this study, which supports the use of cancer registry data in research, demonstrates a potential avenue for improving the cost and efficiency of cancer research in England and Wales. Still, it is important to note that this study only looked at diagnostic data, as SYMPLIFY was a diagnostic study. Thus, although this study found high completeness, concordance, and timeliness of cancer registry data, this can only be said for the diagnostic variables investigated. This study supports the use of cancer data for other diagnostic studies and those concerned with diagnoses, cancer sites, and stages. Thus, epidemiological studies and those concerned with monitoring cancer diagnoses within the population can reliably use cancer registry data in England and Wales.

While many studies are interested in diagnostic data, much of cancer research is concerned with other cancer data, such as treatments, outcomes, and survival. However, the role of registry data in cancer research, as it is concerns treatments and outcomes information, requires further investigation. A previous study found that using cancer registry data for comparative effectiveness research often resulted in survival outcomes that were not concordant with data from randomised clinical trials (51). The authors suggested that possible reasons for these differences observed between randomised trials and comparative effectiveness research include lack of randomisation in clinical effectiveness research, misclassified registry data, the populations used in clinical trials tending to be healthier and having fewer morbidities than the general population of oncology patients, and clinical effectiveness research representing implementation failures rather than treatment failures due to the less technical delivery of a given

treatment (51). Meanwhile, other studies have suggested that differences in cancer registration practices and errors in registration processes may impact estimates of cancer survival (51, 124, 125). Furthermore, some studies have reported higher rates of missing or incorrect data among treatment information in cancer registries (49, 55). Still, variation exists in the accuracy and completion based on treatment types. One study found high accuracy of cancer registry data regarding information on chemotherapy but lower completeness for radiation and hormone therapy (126). Questions regarding the accuracy and completeness of data regarding cancer recurrence following treatment also remain (127). Future work that compares the completeness, accuracy, and timeliness of study data about treatment and outcomes information to that of central registries in England and Wales is necessary to determine if registry data is appropriate for other types of research, which will further help to alleviate the cancer research resource burden.

While this study demonstrates promising findings regarding the completeness, concordance, and timeliness of cancer registry data from England and Wales compared to study data, which supports the use of these registries in diagnostic cancer research, limitations and practicalities remain. For example, in May 2023, there was an issue with data acquisition at one of the English hospital sites, which resulted in a decrease in available cancer data in the May 2023 data cut. The data issue was fixed quickly by the June 2023 data cut. Still, the anomaly of the May 2023 data illustrates the limitations and practical problems that can arise with using routinely collected data in research studies. If the May 2023 data cut had been used for a cancer trial, the number of cancer registrations would have been underestimated due to the missing data, which could drastically impact the study findings. Alternatively, in a study using registry data, the May 2023 data cut would have likely been omitted, which may have affected the ability to

evaluate the study aims. Thus, future studies should allow flexibility in their study timelines in case data issues arise and engage with registries to ensure that any data acquisition issues at sites are communicated early enough to mitigate the impacts of these issues on the study results.

Other practicalities exist regarding he timeliness of data acquisition. In theory, I demonstrated that national cancer registries can support efficient trial delivery that is comparable to data collected on-site. However, adequate resourcing of infrastructure and staff outside of the core national cancer registration process is necessary to ensure the timeliness and quality of data. Other considerations, such as staffing, and the diversion of resources, for example, during the COVID-19 pandemic, can also influence the quality and timeliness of data delivered from cancer registries. To ensure timely data acquisition, national cancer registries should be included early on in trial development and receive adequate funding and resourcing. Engaging with registries early on in research project development and planning will not only improve the likeliness of obtaining sufficient and timely data but will also allow for opportunities to communicate with and learn from the expertise of cancer registry staff. Engaging with, resourcing, and supporting registry staff is crucial to replicate the data quality observed in this study, which was found to be comparable to study data. Even with additional resourcing provided to national cancer registries to support research activities, registry-based research remains a more cost-effective alternative to on-site data collection.

Other limitations must also be considered before engaging in registry-based research. Issues of informed consent, data protection, and patient privacy are important considerations when carrying out research using national cancer registries (8). Ensuring that registry-based research adheres to the same standards of informed consent as standard trials is necessary to uphold ethical standards. Meanwhile, data linkage with cancer registry data provides a practical

challenge. Finally, accessing central registry data is often a concern for researchers. For many researchers, obtaining registry data is complicated by bureaucratic red tape: lengthy forms, paperwork, and the feeling of jumping through hoops to obtain necessary data. To make registry-based research more common and desirable, registries and researchers must work together to make data accessible. Ensuring that data is accessible without unnecessary red tape must be balanced with maintaining a rigorous application process to ensure that patient privacy and data protection are upheld.

5. Conclusion

This study is the first of its kind to comprehensively compare the completeness, concordance, and timeliness of cancer data collected on-site during a prospective cohort study with centrally collected cancer registry data in England and Wales across many cancer sites. Looking broadly across all cancer types, the findings demonstrate that cancer registry data can support diagnostic cancer research through timely, complete, and accurate cancer outcome acquisition. Overall, this study supports the use of cancer registry data to aid study delivery and help alleviate the resource burden associated with cancer research. It also supports calls for adequate cancer registry resourcing to take part in research. However, individual study aims, outcomes, the need for timeliness, and the level of detail required must be considered on a case-by-case basis, as I observed variation between the on-site and registry datasets based on the different data fields and the level of detail investigated.

References

1. Cancer Research UK. Age and Cancer 2021 [cited 07 Sep 2023]. Available from: https://www.cancerresearchuk.org/about-cancer/causes-of-cancer/age-and-cancer#:~:text=1%20in%202%20people%20will,you%20will%20definitely%20get%20cancer.

2. Ahmad AS, Ormiston-Smith N, Sasieni PD. Trends in the lifetime risk of developing cancer in Great Britain: comparison of risk for those born from 1930 to 1960. Br J Cancer. 2015;112(5):943-7.

3. McIntosh SA, Alam F, Adams L, Boon IS, Callaghan J, Conti I, et al. Global funding for cancer research between 2016 and 2020: a content analysis of public and philanthropic investments. Lancet Oncol. 2023;24(6):636-45.

4. Cancer Research UK. How we spend your money [cited 07 Sep 2023]. Available from: https://www.cancerresearchuk.org/about-us/our-organisation/how-we-spend-your-money#:~:text=Our%20annual%20research%20activity,studies%20looking%20at%20cancer%20survivorship).

5. Emanuel EJ, Schnipper LE, Kamin DY, Levinson J, Lichter AS. The costs of conducting clinical research. J Clin Oncol. 2003;21(22):4145-50.

6. What is cancer registration? Cancer Research UK [cited 17 May 2023]. Available from: https://www.cancerresearchuk.org/health-professional/treatment-and-other-post-diagnosis-issues/about-cancer-registration/what-is-cancer-registration#What_is_cancer_registration0.

7. Pop B, Fetica B, Blaga ML, Trifa AP, Achimas-Cadariu P, Vlad CI, et al. The role of medical registries, potential applications and limitations. Med Pharm Rep. 2019;92(1):7-14.

8. Li G, Sajobi TT, Menon BK, Korngut L, Lowerison M, James M, et al. Registry-based randomized controlled trials- what are the advantages, challenges, and areas for future research? J Clin Epidemiol. 2016;80:16-24.

9. Doherty DA, Tong SYC, Reilly J, Shrapnel J, McDonald S, Ahern S, et al. Registry randomised trials: a methodological perspective. BMJ Open. 2023;13(3):e068057.

10. Kwakkenbos L, Imran M, McCall SJ, McCord KA, Fröbert O, Hemkens LG, et al. CONSORT extension for the reporting of randomised controlled trials conducted using cohorts and routinely collected data (CONSORT-ROUTINE): checklist with explanation and elaboration. BMJ. 2021;373:n857.

11. Anderson BR, Gotlieb EG, Hill K, McHugh KE, Scheurer MA, Mery CM, et al. Registry-based trials: a potential model for cost savings? Cardiol Young. 2020;30(6):807-17.

12. Karanatsios B, Prang KH, Verbunt E, Yeung JM, Kelaher M, Gibbs P. Defining key design elements of registry-based randomised controlled trials: a scoping review. Trials. 2020;21(1):552.

13. Rao SV, Hess CN, Barham B, Aberle LH, Anstrom KJ, Patel TB, et al. A registry-based randomized trial comparing radial and femoral approaches in women undergoing percutaneous coronary intervention: the SAFE-PCI for Women (Study of Access Site for Enhancement of PCI for Women) trial. JACC Cardiovasc Interv. 2014;7(8):857-67.

14. Ashrafi R, Hussain H, Brisk R, Boardman L, Weston C. Clinical disease registries in acute myocardial infarction. World J Cardiol. 2014;6(6):415-23.

15. Fröbert O, Lagerqvist B, Olivecrona GK, Omerovic E, Gudnason T, Maeng M, et al. Thrombus aspiration during ST-segment elevation myocardial infarction. N Engl J Med. 2013;369(17):1587-97.

16. Huang SS, Septimus E, Kleinman K, Moody J, Hickok J, Avery TR, et al. Targeted versus universal decolonization to prevent ICU infection. N Engl J Med. 2013;368(24):2255-65.

17. Lauer MS, D'Agostino RB. The randomized registry trial--the next disruptive technology in clinical research? N Engl J Med. 2013;369(17):1579-81.

18. Schwartz AL, Alsan M, Morris AA, Halpern SD. Why Diverse Clinical Trial Participation Matters. N Engl J Med. 2023;388(14):1252-4.

19. El-Galaly TC, Gaidzik VI, Gaman MA, Antic D, Okosun J, Copland M, et al. A Lack of Diversity, Equity, and Inclusion in Clinical Research Has Direct Impact on Patient Care. Hemasphere. 2023;7(3):e842.

20. Shiely F, O Shea N, Murphy E, Eustace J. Registry-based randomised controlled trials: conduct, advantages and challenges-a systematic review. Trials. 2024;25(1):375.

21. McDonald AM, Knight RC, Campbell MK, Entwistle VA, Grant AM, Cook JA, et al. What influences recruitment to randomised controlled trials? A review of trials funded by two UK funding agencies. Trials. 2006;7:9.

22. Carlisle B, Kimmelman J, Ramsay T, MacKinnon N. Unsuccessful trial accrual and human subjects protections: an empirical analysis of recently closed trials. Clin Trials. 2015;12(1):77-83.

23. James S, Rao SV, Granger CB. Registry-based randomized clinical trials--a new clinical trial paradigm. Nat Rev Cardiol. 2015;12(5):312-6.

24. Baer BR, Fremes SE, Gaudino M, Charlson M, Wells MT. On clinical trial fragility due to patients lost to follow up. BMC Med Res Methodol. 2021;21(1):254.

25. Akl EA, Briel M, You JJ, Sun X, Johnston BC, Busse JW, et al. Potential impact on estimated treatment effects of information lost to follow-up in randomised controlled trials (LOST-IT): systematic review. BMJ. 2012;344:e2809.

26. Erlinge D, Omerovic E, Fröbert O, Linder R, Danielewicz M, Hamid M, et al. Bivalirudin versus Heparin Monotherapy in Myocardial Infarction. N Engl J Med. 2017;377(12):1132-42.

27. Götberg M, Christiansen EH, Gudmundsdottir IJ, Sandhall L, Danielewicz M, Jakobsen L, et al. Instantaneous Wave-free Ratio versus Fractional Flow Reserve to Guide PCI. N Engl J Med. 2017;376(19):1813-23.

28. Hofmann R, James SK, Jernberg T, Lindahl B, Erlinge D, Witt N, et al. Oxygen Therapy in Suspected Acute Myocardial Infarction. N Engl J Med. 2017;377(13):1240-9.

29. Holme Ø, Løberg M, Kalager M, Bretthauer M, Hernán MA, Aas E, et al. Long-Term Effectiveness of Sigmoidoscopy Screening on Colorectal Cancer Incidence and Mortality in Women and Men: A Randomized Trial. Ann Intern Med. 2018;168(11):775-82.

30. Bretthauer M, Kaminski MF, Løberg M, Zauber AG, Regula J, Kuipers EJ, et al. Population-Based Colonoscopy Screening for Colorectal Cancer: A Randomized Clinical Trial. JAMA Intern Med. 2016;176(7):894-902.

31. Hall AE, Sanson-Fisher RW, Lynagh MC, Threlfall T, D'Este CA. Format and readability of an enhanced invitation letter did not affect participation rates in a cancer registry-based study: a randomized controlled trial. J Clin Epidemiol. 2013;66(1):85-94.

32. Malila N, Oivanen T, Malminiemi O, Hakama M. Test, episode, and programme sensitivities of screening for colorectal cancer as a public health policy in Finland: experimental design. BMJ. 2008;337:a2261.

33. Thiis-Evensen E, Hoff GS, Sauar J, Langmark F, Majak BM, Vatn MH. Population-based surveillance by colonoscopy: effect on the incidence of colorectal cancer. Telemark Polyp Study I. Scand J Gastroenterol. 1999;34(4):414-20.

34. Auvinen A, Tammela T, Stenman UH, Uusi-Erkkilä I, Leinonen J, Schröder FH, et al. Screening for prostate cancer using serum prostate-specific antigen: a randomised, population-based pilot study in Finland. Br J Cancer. 1996;74(4):568-72.

35. Bohlin KS, Löfgren M, Lindkvist H, Milsom I. Smoking cessation prior to gynecological surgery-A registry-based randomized trial. Acta Obstet Gynecol Scand. 2020;99(9):1230-7.

36. Menon U, Gentry-Maharaj A, Burnell M, Singh N, Ryan A, Karpinskyj C, et al. Ovarian cancer population screening and mortality after long-term follow-up in the UK Collaborative Trial of Ovarian Cancer Screening (UKCTOCS): a randomised controlled trial. Lancet. 2021;397(10290):2182-93.

37. Lensen S, Macnair A, Love SB, Yorke-Edwards V, Noor NM, Martyn M, et al. Access to routinely collected health data for clinical trials - review of successful data requests to UK registries. Trials. 2020;21(1):398.

38. Bray F, Parkin DM. Evaluation of data quality in the cancer registry: principles and methods. Part I: comparability, validity and timeliness. Eur J Cancer. 2009;45(5):747-55.

39. Parkin DM, Bray F. Evaluation of data quality in the cancer registry: principles and methods Part II. Completeness. Eur J Cancer. 2009;45(5):756-64.

40. Gunnarsson U, Seligsohn E, Jestin P, Påhlman L. Registration and validity of surgical complications in colorectal cancer surgery. Br J Surg. 2003;90(4):454-9.

41. Gunnarsson U. Quality assurance in surgical oncology. Colorectal cancer as an example. Eur J Surg Oncol. 2003;29(1):89-94.

42. Himmelman JG, Merrimen J, Matheson K, Theriault C, Wood LA. Accuracy of kidney cancer diagnosis and histological subtype within Canadian cancer registry data. Can Urol Assoc J. 2017;11(9):E326-E9.

43. Holland P, Karmas E, Merrimen J, Wood LA. Accuracy of germ cell tumor histology and stage within a Canadian cancer registry. Can Urol Assoc J. 2023;17(2):44-8.

44. Seneviratne S, Campbell I, Scott N, Shirley R, Peni T, Lawrenson R. Accuracy and completeness of the New Zealand Cancer Registry for staging of invasive breast cancer. Cancer Epidemiol. 2014;38(5):638-44.

45. Stevens W, Stevens G, Kolbe J, Cox B. Comparison of New Zealand Cancer Registry data with an independent lung cancer audit. N Z Med J. 2008;121(1276):29-41.

46. Cunningham R, Sarfati D, Hill S, Kenwright D. An audit of colon cancer data on the New Zealand Cancer Registry. N Z Med J. 2008;121(1279):46-56.

47. Ballantine KR, Hanna S, Macfarlane S, Bradbeer P, Teague L, Hunter S, et al. Childhood cancer registration in New Zealand: A registry collaboration to assess and improve data quality. Cancer Epidemiol. 2018;55:104-9.

48. Thoburn KK, German RR, Lewis M, Nichols PJ, Ahmed F, Jackson-Thompson J. Case completeness and data accuracy in the Centers for Disease Control and Prevention's National Program of Cancer Registries. Cancer. 2007;109(8):1607-16.

49. Traverso-Ortiz M, Duran D, Mesnard M, Ng D, Dailey S. Results of Year 2 Data Quality Evaluation of CDC's National Program of Cancer Registries: Weighing the Evidence, Identifying Research Gaps, and Evaluating Outputs of a Prevention Research Agenda. J Registry Manag. 2022;49(2):75-8.

50. Winchester DP, Stewart AK, Phillips JL, Ward EE. The national cancer data base: past, present, and future. Ann Surg Oncol. 2010;17(1):4-7.

51. Kumar A, Guss ZD, Courtney PT, Nalawade V, Sheridan P, Sarkar RR, et al. Evaluation of the Use of Cancer Registry Data for Comparative Effectiveness Research. JAMA Netw Open. 2020;3(7):e2011985.

52. Søgaard M, Olsen M. Quality of cancer registry data: completeness of TNM staging and potential implications. Clin Epidemiol. 2012;4 Suppl 2(Suppl 2):1-3.

53. Jensen AR, Overgaard J, Storm HH. Validity of breast cancer in the Danish Cancer Registry. A study based on clinical records from one county in Denmark. Eur J Cancer Prev. 2002;11(4):359-64.

54. Ingimarsdóttir IJ, Rusch E, Engholm G, Storm HH, Brasso K. Quality assessment of prostate cancer reports to the Danish Cancer Registry. Acta Oncol. 2016;55(1):24-9.

55. Lunkka P, Malila N, Ryynänen H, Heikkinen S, Sallinen V, Koskenvuo L. Accuracy of Finnish Cancer Registry colorectal cancer data: a comparison between registry data and clinical records. Scand J Gastroenterol. 2021;56(3):247-51.

56. Korhonen P, Malila N, Pukkala E, Teppo L, Albanes D, Virtamo J. The Finnish Cancer Registry as follow-up source of a large trial cohort--accuracy and delay. Acta Oncol. 2002;41(4):381-8.

57. Larsen IK, Småstuen M, Johannesen TB, Langmark F, Parkin DM, Bray F, et al. Data quality at the Cancer Registry of Norway: an overview of comparability, completeness, validity and timeliness. Eur J Cancer. 2009;45(7):1218-31.

58. Ramos M, Franch P, Zaforteza M, Artero J, Durán M. Completeness of T, N, M and stage grouping for all cancers in the Mallorca Cancer Registry. BMC Cancer. 2015;15:847.

59. Cero MD, Rodríguez-Santiago J, Miró M, Castro S, Miranda C, Santamaría M, et al. Evaluation of data quality in the Spanish EURECCA Esophagogastric Cancer Registry. Eur J Surg Oncol. 2021;47(12):3081-7.

60. Moberger P, Sköldberg F, Birgisson H. Evaluation of the Swedish Colorectal Cancer Registry: an overview of completeness, timeliness, comparability and validity. Acta Oncol. 2018;57(12):1611-21.

61. Jörgren F, Johansson R, Damber L, Lindmark G. Validity of the Swedish Rectal Cancer Registry for patients treated with major abdominal surgery between 1995 and 1997. Acta Oncol. 2013;52(8):1707-14.

62. Linder G, Lindblad M, Djerf P, Elbe P, Johansson J, Lundell L, et al. Validation of data quality in the Swedish National Register for Oesophageal and Gastric Cancer. Br J Surg. 2016;103(10):1326-35.

63. Tomic K, Sandin F, Wigertz A, Robinson D, Lambe M, Stattin P. Evaluation of data quality in the National Prostate Cancer Register of Sweden. Eur J Cancer. 2015;51(1):101-11.

64. Barlow L, Westergren K, Holmberg L, Talbäck M. The completeness of the Swedish Cancer Register: a sample survey for year 1998. Acta Oncol. 2009;48(1):27-33.

65. Löfgren L, Eloranta S, Krawiec K, Asterkvist A, Lönnqvist C, Sandelin K, et al. Validation of data quality in the Swedish National Register for Breast Cancer. BMC Public Health. 2019;19(1):495.

66. Holmäng S, Amsler-Nordin S, Carlson K, Holmberg E, Johansson SL. Completeness and correctness of registration of renal pelvic and ureteral cancer in the Swedish Cancer Registry. Scand J Urol Nephrol. 2008;42(1):12-7.

67. Kearney TM, Donnelly C, Kelly JM, O'Callaghan EP, Fox CR, Gavin AT. Validation of the completeness and accuracy of the Northern Ireland Cancer Registry. Cancer Epidemiol. 2015;39(3):401-4.

68. Brewster D, Crichton J, Muir C. How accurate are Scottish cancer registration data? Br J Cancer. 1994;70(5):954-9.

69. Brewster D, Muir C, Crichton J. Registration of lung cancer in Scotland: an assessment of data accuracy based on review of medical records. Cancer Causes Control. 1995;6(4):303-10.

70. Henson KE, Elliss-Brookes L, Coupland VH, Payne E, Vernon S, Rous B, et al. Data Resource Profile: National Cancer Registration Dataset in England. Int J Epidemiol. 2020;49(1):16-h.

71. National Cancer Registration and Analysis Service. Rapid Cancer Registration Dataset [Cited 20 Sep 2023]. Available from: http://www.ncin.org.uk/collecting_and_using_data/rcrd.

72. NHS Digital. Rapid Cancer Registration Data. [Cited 20 Sep 2023]. Available from: https://digital.nhs.uk/data-and-information/publications/statistical/mi-rapid-cancer-registration-data/current/current#:~:text=The%20rapid%20cancer%20registration%20dataset,NHS%20Digital%20and%20other%20agencies.

73. NDRS. COVID-19 rapid cancer registration and treatment data 2023. [Cited 11 Apr 2024]. Available from: https://digital.nhs.uk/ndrs/data/data-outputs/covid-19-rcrd-and-treatment-data.

74. NDRS. Using RCRD and staging data 2023. [Cited 12 Mar 2024]. Available from: https://digital.nhs.uk/ndrs/data/data-sets/rcrd/using-rcrd-to-measure-stage.

75. Portal EHI. National Cancer Registry for Wales 2023. [Cited 12 Mar 2024]. Available from: https://www.healthinformationportal.eu/health-information-sources/national-cancer-registry-wales.

76. Wales PH. Welsh Cancer Intelligence and Surveillance Unit (WCISU) - About Us. [Cited 12 Mar 2024]. Available from: https://phw.nhs.wales/services-and-teams/welsh-cancer-intelligence-and-surveillance-unit-wcisu/about-us/.

77. Digital Health and Care Wales. About Digital Health and Care Wales. [Cited 22 Jun 2024]. Available from: https://dhcw.nhs.wales/about-us/.

78. Jones AM, Morris E, Thomas J, Forman D, Melia J, Moss SM. Evaluation of bowel cancer registration data in England, 1996-2004. Br J Cancer. 2009;101(8):1269-73.

79. West Midlands Cancer Intelligence Unit. Quantifying the Completeness of National Breast Cancer Data (cases diagnosed in 2006) Executive Summary. National Cancer Intelligence Network; 2009.

80. Merriel SWD, Turner EL, Walsh E, Young GJ, Metcalfe C, Hounsome L, et al. Cross-sectional study evaluating data quality of the National Cancer Registration and Analysis Service (NCRAS) prostate cancer registry data using the Cluster randomised trial of PSA testing for Prostate cancer (CAP). BMJ Open. 2017;7(11):e015994.

81. UKIACR. Performance Indicators Commentaries for Tumours Diagnosed in 2019.; 2021.

82. Nicholson BD, Oke J, Virdee PS, Harris DA, O'Doherty C, Park JE, et al. Multi-cancer early detection test in symptomatic patients referred for cancer investigation in England and Wales (SYMPLIFY): a large-scale, observational cohort study. Lancet Oncol. 2023;24(7):733-43.

83. Johns Hopkins. Understanding ICD-10 [cited 2024 13 May]. Available from: https://www.hopkinsmedicine.org/johns-hopkins-health-plans/providers-physicians/icd-10.

84. World Health Organization. International statistical classification of diseases and health related problems 10th revision. 2015.

85. World Health Organization. International Classification of Diseases for Oncology, 3rd Edition (ICD-O-3) [cited 2024 13 May]. Available from: https://www.who.int/standards/classifications/other-classifications/international-classification-of-diseases-for-oncology.

86. Federal Institute for Drugs and Medical Devices. ICD-O-3. [Cited 20 Sep 2023]. Available from: https://www.bfarm.de/EN/Code-systems/Classifications/ICD/ICD-O-3/_node.html#:~:text=Robert%2DKoch%2DInstitut-,Structure%20of%20ICD%2DO%20%2D3,ICD%20%2D10%20for%20malignant%20neoplasms.

87. World Health Organization. International Classification of Diseases for Oncology, Third Edition, First Revision. 2013.

88. Dickinson HO, Salotti JA, Birch PJ, Reid MM, Malcolm A, Parker L. How complete and accurate are cancer registrations notified by the National Health Service Central Register for England and Wales? J Epidemiol Community Health. 2001;55(6):414-22.

89. McPhail S, Johnson S, Greenberg D, Peake M, Rous B. Stage at diagnosis and early mortality from cancer in England. Br J Cancer. 2015;112 Suppl 1(Suppl 1):S108-15.

90. Hoopes M, Voss R, Angier H, Marino M, Schmidt T, DeVoe JE, et al. Assessing Cancer History Accuracy in Primary Care Electronic Health Records Through Cancer Registry Linkage. J Natl Cancer Inst. 2021;113(7):924-32.

91. Cancer Research UK. TNM Staging | Bowel Cancer 2022. [Cited 11 Apr 2024]. Available from: https://www.cancerresearchuk.org/about-cancer/bowel-cancer/stages-types-and-grades/TNM-staging.

92. Cancer Research UK. TNM staging for pancreatic cancer 2023 [Cited 11 Apr 2024]. Available from: https://www.cancerresearchuk.org/about-cancer/pancreatic-cancer/stages-types-grades/tnm-staging.

93. NDRS. Cancer 2023. [Cited 11 Apr 2024]. Available from: https://digital.nhs.uk/ndrs/about/ncras.

94. Ramón Y Cajal S, Sesé M, Capdevila C, Aasen T, De Mattos-Arruda L, Diaz-Cano SJ, et al. Clinical implications of intratumor heterogeneity: challenges and opportunities. J Mol Med (Berl). 2020;98(2):161-77.

95. Komatsu S, Ichikawa D, Miyamae M, Kosuga T, Konishi H, Shiozaki A, et al. Discrepancies in the histologic type between biopsy and resected specimens: a cautionary note for mixed-type gastric carcinoma. World J Gastroenterol. 2015;21(15):4673-9.

96. Kim Y, Yoon HJ, Kim JH, Chun J, Youn YH, Park H, et al. Effect of histologic differences between biopsy and final resection on treatment outcomes in early gastric cancer. Surg Endosc. 2020;34(11):5046-54.

97. Butler J, Foot C, Bomb M, Hiom S, Coleman M, Bryant H, et al. The International Cancer Benchmarking Partnership: an international collaboration to inform cancer policy in Australia, Canada, Denmark, Norway, Sweden and the United Kingdom. Health Policy. 2013;112(1-2):148-55.

98. Rosen RD, Sapra A. TNM Classification. StatPearls Publishing; 2024.

99. Leukemia & Lymphoma Society. CLL Staging. [Cited 15 Apr 2024]. Available from: https://www.lls.org/leukemia/chronic-lymphocytic-leukemia/diagnosis/cll-staging.

100. NHS North Central London Cancer Alliance. Rapid Diagnostic Centres. [Cited 15 Apr 2024]. Available from: https://www.nclcanceralliance.nhs.uk/our-work/diagnosis-and-treatment/rapid-diagnostic-centres/.

101. Blood cancer UK. Blood cancer symptoms and signs. [Cited 15 Apr 2024]. Available from: https://bloodcancer.org.uk/understanding-blood-cancer/blood-cancer-signs-symptoms/#blood-cancer-symptoms.

102. Canadian Cancer Society. Staging cancer. [Cited 15 Apr 2024]. Available from: https://cancer.ca/en/cancer-information/what-is-cancer/stage-and-grade/staging.

103. American Cancer Society. Cancer Staging 2022. [Cited 15 Apr 2024]. Available from: https://www.cancer.org/cancer/diagnosis-staging/staging.html.

104. McCluggage WG, Hirschowitz L, Ganesan R, Kehoe S, Nordin A. Which staging system to use for gynaecological cancers: a survey with recommendations for practice in the UK. J Clin Pathol. 2010;63(9):768-70.

105. Brierley J, Gospodarowicz M, O'Sullivan B. The principles of cancer staging. Ecancermedicalscience. 2016;10:ed61.

106. Matsunaga T, Suzuki K, Hattori A, Fukui M, Hayashi T, Takamochi K. A problem with clinical T factor in the 8th TNM edition: Prognosis and EGFR mutation status of small sized lung cancers with difficulty to measure the diameter of solid component in part-solid tumor. Lung Cancer. 2023;184:107354.

107. Shia J, Klimstra DS, Bagci P, Basturk O, Adsay NV. TNM staging of colorectal carcinoma: issues and caveats. Semin Diagn Pathol. 2012;29(3):142-53.

108. Ganeshalingam S, Koh DM. Nodal staging. Cancer Imaging. 2009;9(1):104-11.

109. Rashid RJ, Tahir SH, Kakamad FH, Omar SS, Salih AM, Ahmed SF, et al. Whole-body MRI for metastatic workup in patients diagnosed with cancer. Mol Clin Oncol. 2023;18(4):33.

110. Cancer Research UK. Lung cancer statistics. [Cited 16 Apr 2024]. Available from: https://www.cancerresearchuk.org/health-professional/cancer-statistics/statistics-by-cancer-type/lung-cancer#heading-One.

111. Di Girolamo C, Walters S, Benitez Majano S, Rachet B, Coleman MP, Njagi EN, et al. Characteristics of patients with missing information on stage: a population-based study of patients diagnosed with colon, lung or breast cancer in England in 2013. BMC Cancer. 2018;18(1):492.

112. NHS England. Earlier and faster diagnosis. [Cited 3 May 2024]. Available from: https://www.england.nhs.uk/cancer/quarterly-report-overviews/q4-2021-q1-2122/earlier-and-faster-diagnosis/#:~:text=The%20Faster%20Diagnosis%20Standard%20is,quarters%20(75%25)%20of%20patients.

113. European Network of Cancer Registries. Coding Incidence Date. 2022.

114. Coleman MP. Multiple primary malignancy in England and Wales, 1971-80: a pilot study of OPCS records. Community Med. 1987;9(1):15-24.

115. Merriman KW, Broome RG, De Las Pozas G, Landvogt LD, Qi Y, Keating J. Evolution of the Cancer Registrar in the Era of Informatics. JCO Clin Cancer Inform. 2021;5:272-8.

116. Zanetti R, Schmidtmann I, Sacchetto L, Binder-Foucard F, Bordoni A, Coza D, et al. Completeness and timeliness: Cancer registries could/should improve their performance. Eur J Cancer. 2015;51(9):1091-8.

117. Macmillan Cancer Support. Rare Cancers. [Cited 23 Jun 2024]. Available from: https://www.macmillan.org.uk/cancer-information-and-support/rare-cancers.

118. Cancer Research UK. Cancer Statistics for the UK. [Cited 15 Aug 2024]. Available from: https://www.cancerresearchuk.org/health-professional/cancer-statistics-for-the-uk#heading-Four.

119. Dang A. Real-World Evidence: A Primer. Pharmaceut Med. 2023;37(1):25-36.

120. Chodankar D. Introduction to real-world evidence studies. Perspect Clin Res. 2021;12(3):171-4.

121. Patsopoulos NA. A pragmatic view on pragmatic trials. Dialogues Clin Neurosci. 2011;13(2):217-24.

122. UK Clinical Research Collaboration 2023. UK Health Research Analysis 2022. 2024.

123. Cancer Research UK. Longer, better lives A manifesto for cancer research and care. 2023.

124. Andersson TM, Rutherford MJ, Myklebust T, Møller B, Soerjomataram I, Arnold M, et al. Exploring the impact of cancer registry completeness on international cancer survival differences: a simulation study. Br J Cancer. 2021;124(5):1026-32.

125. Rutherford MJ, Møller H, Lambert PC. A comprehensive assessment of the impact of errors in the cancer registration process on 1- and 5-year relative survival estimates. Br J Cancer. 2013;108(3):691-8.

126. Anderson C, Baggett CD, Rao C, Moy L, Kushi LH, Chao CR, et al. Validity of state cancer registry treatment information for adolescent and young adult women. Cancer Epidemiol. 2020;64:101652.

127. Sutton EA, Kamdem Talom BC, Ebner DK, Weiskittel TM, Breen WG, Kowalchuk RO, et al. Accuracy of a Cancer Registry Versus Clinical Care Team Chart Abstraction in Identifying Cancer Recurrence. Mayo Clin Proc Innov Qual Outcomes. 2024;8(3):225-31.

Appendix

Table 1. ICD-O-3 broad morphology groupings.

ICD-O-3 Code	Morphology	Broad Morphology Grouping
8000	Neoplasm	Neoplasms, NOS
8001	Tumour cells	Neoplasms, NOS
8002	Malignant tumour, small cell type	Neoplasms, NOS
8003	Malignant tumour, giant cell type	Neoplasms, NOS
8004	Malignant tumour, spindle cell type	Neoplasms, NOS
8005	Malignant tumour, clear cell type	Neoplasms, NOS
8010	Carcinoma, NOS	Epithelial neoplasms, NOS
8011	Epithelioma, malignant	Epithelial neoplasms, NOS
8012	Large cell carcinoma, NOS	Epithelial neoplasms, NOS
8013	Large call neuroendocrine carcinoma	Epithelial neoplasms, NOS
8014	Large cell carcinoma with rhabdoid phenotype	Epithelial neoplasms, NOS
8015	Glassy cell carcinoma	Epithelial neoplasms, NOS
8020	Carcinoma, undifferentiated, NOS	Epithelial neoplasms, NOS
8021	Carcinoma, anaplastic, NOS	Epithelial neoplasms, NOS
8022	Pleomorphic carcinoma	Epithelial neoplasms, NOS
8030	Giant cell and spindle cell carcinoma	Epithelial neoplasms, NOS
8031	Giant cell carcinoma	Epithelial neoplasms, NOS
8032	Spindle cell carcinoma, NOS	Epithelial neoplasms, NOS
8033	Pseudosarcomatous carcinoma	Epithelial neoplasms, NOS
8034	Polygonal cell carcinoma	Epithelial neoplasms, NOS
8035	Carcinoma with osteoclast-like giant cells	Epithelial neoplasms, NOS
8041	Small cell carcinoma, NOS	Epithelial neoplasms, NOS

8042	Oat cell carcinoma	Epithelial neoplasms, NOS
8043	Small cell carcinoma, fusiform cell	Epithelial neoplasms, NOS
8044	Small cell carcinoma, intermediate cell	Epithelial neoplasms, NOS
8045	Combined small cell carcinoma	Epithelial neoplasms, NOS
8046	Non-small cell carcinoma	Epithelial neoplasms, NOS
8050	Papillary carcinoma, NOS	Squamous cell neoplasms
8051	Verrucous carcinoma, NOS	Squamous cell neoplasms
8052	Papillary squamous cell carcinoma	Squamous cell neoplasms
8070	Squamous cell carcinoma, NOS	Squamous cell neoplasms
8071	Squamous cell carcinoma, keratinizing, NOS	Squamous cell neoplasms
8072	Squamous cell carcinoma, large cell, nonkeratinizing, NOS	Squamous cell neoplasms
8073	Squamous cell carcinoma, small cell, nonkeratinizing	Squamous cell neoplasms
8074	Squamous cell carcinoma, spindle cell	Squamous cell neoplasms
8075	Squamous cell carcinoma, adenoid	Squamous cell neoplasms
8076	Squamous cell carcinoma, microinvasive	Squamous cell neoplasms
8078	Squamous cell carcinoma with horn formation	Squamous cell neoplasms
8082	Lymphoepithelial carcinoma	Squamous cell neoplasms
8083	Basaloid squamous cell carcinoma	Squamous cell neoplasms
8084	Squamous cell carcinoma, clear cell type	Squamous cell neoplasms
8090	Basal cell carcinoma, NOS (C44._)	Basal cell neoplasms
8091	Multifocal superficial basal cell carcinoma (C44._)	Basal cell neoplasms
8092	Infiltrating basal cell carcinoma, NOS (C44._)	Basal cell neoplasms
8093	Basal cell carcinoma, fibroepithelial (C44._)	Basal cell neoplasms
8094	Basosquamous carcinoma (C44._)	Basal cell neoplasms
8095	Metatypical carcinoma (C44._)	Basal cell neoplasms
8097	Basal cell carcinoma, nodular (C44._)	Basal cell neoplasms
8098	Adenoid basal carcinoma (C53._)	Basal cell neoplasms
8102	Trichilemmocarcinoma (C44._)	Basal cell neoplasms

8110	Pilomatrix carcinoma (C44._)	Basal cell neoplasms
8120	Transitional cell carcinoma, NOS	Transitional cell papillomas and carcinomas
8121	Schneiderian carcinoma (C30.0, C31._)	Transitional cell papillomas and carcinomas
8122	Transitional cell carcinoma, spindle cell	Transitional cell papillomas and carcinomas
8123	Basaloid carcinoma	Transitional cell papillomas and carcinomas
8124	Cloacogenic carcinoma (C21.2)	Transitional cell papillomas and carcinomas
8130	Papillary transitional cell carcinoma (C67._)	Transitional cell papillomas and carcinomas
8131	Transitional cell carcinoma, micropapillary (C67._)	Transitional cell papillomas and carcinomas
8140	Adenocarcinoma, NOS	Adenomas and adenocarcinomas
8141	Scirrhous adenocarcinoma	Adenomas and adenocarcinomas
8142	Linitis plastica (C16._)	Adenomas and adenocarcinomas
8143	Superficial spreading adenocarcinoma	Adenomas and adenocarcinomas
8144	Adenocarcinoma, intestinal type (C16._)	Adenomas and adenocarcinomas
8145	Carcinoma, diffuse type (C16._)	Adenomas and adenocarcinomas
8147	Vasal cell adenocarcinoma	Adenomas and adenocarcinomas
8150	Pancreatic endocrine tumour (C25._)	Adenomas and adenocarcinomas
8151	Insulinoma, malignant (C25._)	Adenomas and adenocarcinomas
8152	Glucagonoma, malignant (C25._)	Adenomas and adenocarcinomas
8153	Gastrinoma, malignant	Adenomas and adenocarcinomas
8154	Mixed pancreatic endocrine and exocrine tumour, malignant (C25._)	Adenomas and adenocarcinomas
8155	Vipoma	Adenomas and adenocarcinomas
8156	Somatostatinoma	Adenomas and adenocarcinomas
8160	Cholangiocarcinoma (C22.1, C24.0)	Adenomas and adenocarcinomas
8161	Bile duct cystadenocarcinoma (C22.1, C24.0)	Adenomas and adenocarcinomas
8162	Klatskin tumour (C22.1, C24.0)	Adenomas and adenocarcinomas
8163	Pancreatobiliary-type carcinoma (C24.1)	Adenomas and adenocarcinomas
8170	Hepatocellular carcinoma, NOS (C22.0)	Adenomas and adenocarcinomas
8171	Hepatocellular carcinoma, fibrolamellar (C22.0)	Adenomas and adenocarcinomas

8172	Hepatocellular carcinoma, scirrhous (C22.0)	Adenomas and adenocarcinomas
8173	Hepatocellular carcinoma, spindle cell variant (C22.0)	Adenomas and adenocarcinomas
8174	Hepatocellular carcinoma, clear cell type (C22.0)	Adenomas and adenocarcinomas
8175	Hepatocellular carcinoma, pleomorphic type (C22.0)	Adenomas and adenocarcinomas
8180	Combined hepatocellular carcinoma and cholangiocarcinoma (C22.0)	Adenomas and adenocarcinomas
8190	Trabecular adenocarcinoma	Adenomas and adenocarcinomas
8200	Adenoid cystic carcinoma	Adenomas and adenocarcinomas
8201	Cribriform carcinoma, NOS	Adenomas and adenocarcinomas
8210	Adenocarcinoma in adenomatous polyp	Adenomas and adenocarcinomas
8211	Tubular adenocarcinoma	Adenomas and adenocarcinomas
8213	Serrated adenocarcinoma	Adenomas and adenocarcinomas
8214	Parietal cell carcinoma (C16._)	Adenomas and adenocarcinomas
8215	Adenocarcinoma of anal glands (C21.1)	Adenomas and adenocarcinomas
8220	Adenocarcinoma in adenomatous polyposis coli (C18._)	Adenomas and adenocarcinomas
8221	Adenocarcinoma in multiple adenomatous polyps (C18._)	Adenomas and adenocarcinomas
8230	Solid carcinoma, NOS	Adenomas and adenocarcinomas
8231	Carcinoma simplex	Adenomas and adenocarcinomas
8240	Carcinoid tumour, NOS	Adenomas and adenocarcinomas
8241	Enterochromaffin cell carcinoid	Adenomas and adenocarcinomas
8242	Enterochromaffin-like cell tumour	Adenomas and adenocarcinomas
8243	Goblet cell carcinoid	Adenomas and adenocarcinomas
8244	Mixed adenoneuroendocrine carcinoma	Adenomas and adenocarcinomas
8245	Adenocarcinoid tumour	Adenomas and adenocarcinomas
8246	Neuroendocrine carcinoma, NOS	Adenomas and adenocarcinomas
8247	Merkel cell carcinoma (C44._)	Adenomas and adenocarcinomas
8249	Atypical carcinoid tumour	Adenomas and adenocarcinomas
8250	Bronchiolo-alveolar adenocarcinoma, NOS	Adenomas and adenocarcinomas
8251	Alveolar adenocarcinoma (C34._)	Adenomas and adenocarcinomas
8252	Bronchiolo-alveolar carcinoma, non-mucinous (C34._)	Adenomas and adenocarcinomas

8253	Bronchiolo-alveolar carcinoma, mucinous (C34._)	Adenomas and adenocarcinomas
8254	Bronchiolo-alveolar carcinoma, mixed mucinous and non-mucinous (C34._)	Adenomas and adenocarcinomas
8255	Adenocarcinoma with mixed subtypes	Adenomas and adenocarcinomas
8260	Papillary adenocarcinoma, NOS	Adenomas and adenocarcinomas
8261	Adenocarcinoma in villous adenoma	Adenomas and adenocarcinomas
8262	Villous adenocarcinoma	Adenomas and adenocarcinomas
8263	Adenocarcinoma in tubulovillous adenoma	Adenomas and adenocarcinomas
8265	Micropapillary carcinoma, NOS (C18._, C19.9, C20.9)	Adenomas and adenocarcinomas
8270	Chromophobe carcinoma (C75.1)	Adenomas and adenocarcinomas
8272	Pituitary carcinoma, NOS (C75.1)	Adenomas and adenocarcinomas
8280	Acidophil carcinoma (C75.1)	Adenomas and adenocarcinomas
8281	Mixed acidophil-basophil carcinoma (C75.1)	Adenomas and adenocarcinomas
8290	Oxyphilic adenocarcinoma	Adenomas and adenocarcinomas
8300	Basophil carcinoma (C75.1)	Adenomas and adenocarcinomas
8310	Clear cell adenocarcinoma, NOS	Adenomas and adenocarcinomas
8312	Renal cell carcinoma, NOS (C64.9)	Adenomas and adenocarcinomas
8313	Clear cell adenocarcinofibroma (C56.9)	Adenomas and adenocarcinomas
8314	Lipid-rich carcinoma (C50._)	Adenomas and adenocarcinomas
8315	Glycogen-rich carcinoma	Adenomas and adenocarcinomas
8316	Cyst-associated renal cell carcinoma (C64.9)	Adenomas and adenocarcinomas
8317	Renal cell carcinoma, chromophobe type (C64.9)	Adenomas and adenocarcinomas
8318	Renal cell carcinoma, sarcomatoid (C64.9)	Adenomas and adenocarcinomas
8319	Collecting duct carcinoma (C64.9)	Adenomas and adenocarcinomas
8320	Glandular cell carcinoma	Adenomas and adenocarcinomas
8322	Water-clear cell adenocarcinoma (C75.0)	Adenomas and adenocarcinomas
8323	Mixed cell adenocarcinoma	Adenomas and adenocarcinomas
8330	Follicular adenocarcinoma, NOS (C73.9)	Adenomas and adenocarcinomas
8331	Follicular adenocarcinoma, well differentiated (C73.9)	Adenomas and adenocarcinomas
8332	Follicular adenocarcinoma, trabecular (C73.9)	Adenomas and adenocarcinomas

8333	Fetal adenocarcinoma	Adenomas and adenocarcinomas
8335	Follicular carcinoma, minimally invasive (C73.9)	Adenomas and adenocarcinomas
8337	Insular carcinoma (C73.9)	Adenomas and adenocarcinomas
8340	Papillary carcinoma, follicular variant (C73.9)	Adenomas and adenocarcinomas
8341	Papillary microcarcinoma (C73.9)	Adenomas and adenocarcinomas
8342	Papillary carcinoma, oxyphilic cell (C73.9)	Adenomas and adenocarcinomas
8343	Papillary carcinoma, encapsulated (C73.9)	Adenomas and adenocarcinomas
8344	Papillary carcinoma, columnar cell (C73.9)	Adenomas and adenocarcinomas
8345	Medullary carcinoma with amyloid stroma (C73.9)	Adenomas and adenocarcinomas
8346	Mixed medullary-follicular carcinoma (C73.9)	Adenomas and adenocarcinomas
8347	Mixed medullary-papillary carcinoma (C73.9)	Adenomas and adenocarcinomas
8350	Nonencapsulated sclerosing carcinoma (C73.9)	Adenomas and adenocarcinomas
8370	Adrenal cortical carcinoma (C74.0)	Adenomas and adenocarcinomas
8380	Endometrioid adenocarcinoma, NOS	Adenomas and adenocarcinomas
8381	Endometrioid adenofibroma, malignant	Adenomas and adenocarcinomas
8382	Endometrioid adenocarcinoma, secretory variant	Adenomas and adenocarcinomas
8383	Endometrioid adenocarcinoma, ciliated cell variant	Adenomas and adenocarcinomas
8384	Adenocarcinoma, endocervical type	Adenomas and adenocarcinomas
8390	Skin appendage carcinoma (C44._)	Adnexal and skin appendage neoplasms
8400	Sweat gland adenocarcinoma (C44._)	Adnexal and skin appendage neoplasms
8401	Apocrine adenocarcinoma (C44._)	Adnexal and skin appendage neoplasms
8402	Nodular hidradenoma, malignant (C44._)	Adnexal and skin appendage neoplasms
8403	Malignant eccrine spiradenoma (C44._)	Adnexal and skin appendage neoplasms
8407	Sclerosing sweat duct carcinoma (C44._)	Adnexal and skin appendage neoplasms
8408	Eccrine papillary adenocarcinoma (C44._)	Adnexal and skin appendage neoplasms
8409	Eccrine poroma, malignant	Adnexal and skin appendage neoplasms
8410	Sebaceous adenocarcinoma (C44._)	Adnexal and skin appendage neoplasms
8413	Eccrine adenocarcinoma (C44._)	Adnexal and skin appendage neoplasms

8420	Ceruminous adenocarcinoma (C44.2)	Adnexal and skin appendage neoplasms
8430	Mucoepidermoid carcinoma	Mucoepidermoid neoplasms
8440	Cystadenocarcinoma	Cystic, mucinous, and serous neoplasms
8441	Serous cystadenocarcinoma, NOS (C56.9)	Cystic, mucinous, and serous neoplasms
8450	Papillary cystadenocarcinoma, NOS (C56.9)	Cystic, mucinous, and serous neoplasms
8452	Solid pseudopapillary carcinoma (C25._)	Cystic, mucinous, and serous neoplasms
8453	Intraductal papillary mucinous carcinoma, invasive (C25._)	Cystic, mucinous, and serous neoplasms
8460	Papillary serous cystadenocarcinoma (C56.9)	Cystic, mucinous, and serous neoplasms
8461	Serous surface papillary carcinoma (C56.9)	Cystic, mucinous, and serous neoplasms
8470	Mucinous cystadenocarcinoma, NOS (C56.9)	Cystic, mucinous, and serous neoplasms
8471	Papillary mucinous cystadenocarcinoma, NOS (C56.9)	Cystic, mucinous, and serous neoplasms
8480	Mucinous adenocarcinoma	Cystic, mucinous, and serous neoplasms
8481	Mucin-producing adenocarcinoma	Cystic, mucinous, and serous neoplasms
8482	Mucinous adenocarcinoma endocervical type	Cystic, mucinous, and serous neoplasms
8490	Signet ring cell carcinoma	Cystic, mucinous, and serous neoplasms
8500	Infiltrating duct carcinoma, NOS (C50._)	Ductal and lobular neoplasms
8501	Comedocarcinoma, NOS (C50._)	Ductal and lobular neoplasms
8502	Secretory carcinoma of breast (C50._)	Ductal and lobular neoplasms
8503	Intraductal papillary adenocarcinoma with invasion (C50._)	Ductal and lobular neoplasms
8504	Intracystic carcinoma, NOS	Ductal and lobular neoplasms
8508	Cystic hypersecretory carcinoma (C50._)	Ductal and lobular neoplasms
8510	Medullary carcinoma, NOS	Ductal and lobular neoplasms
8512	Medullary carcinoma with lymphoid stroma	Ductal and lobular neoplasms
8513	Atypical medullary carcinoma (C50._)	Ductal and lobular neoplasms
8514	Duct carcinoma, desmoplastic type	Ductal and lobular neoplasms
8520	Lobular carcinoma, NOS (C50._)	Ductal and lobular neoplasms
8521	Infiltrating ductular carcinoma (C50._)	Ductal and lobular neoplasms
8522	Infiltrating duct and lobular carcinoma (C50._)	Ductal and lobular neoplasms

8523	Infiltrating duct mixed with other types of carcinoma (C50._)	Ductal and lobular neoplasms
8524	Infiltrating lobular mixed with other types of carcinoma (C50._)	Ductal and lobular neoplasms
8525	Polymorphous low grade adenocarcinoma	Ductal and lobular neoplasms
8530	Inflammatory carcinoma (C50._)	Ductal and lobular neoplasms
8540	Paget disease, mammary(C50._)	Ductal and lobular neoplasms
8541	Paget disease, and infiltrating duct carcinoma of breast (C50._)	Ductal and lobular neoplasms
8542	Paget disease, extramammary (except Paget disease of bone)	Ductal and lobular neoplasms
8543	Paget disease and intraductal carcinoma of breast (C50._)	Ductal and lobular neoplasms
8550	Acinar cell carcinoma	Acinar cell neoplasms
8551	Acinar cell cystadenocarcinoma	Acinar cell neoplasms
8552	Mixed acinar-ductal carcinoma	Acinar cell neoplasms
8560	Adenosquamous carcinoma	Complex epithelial neoplasms
8562	Epithelial-myoepithelial carcinoma	Complex epithelial neoplasms
8570	Adenocarcinoma with squamous metaplasia	Complex epithelial neoplasms
8571	Adenocarcinoma with cartilaginous and osseous metaplasia	Complex epithelial neoplasms
8572	Adenocarcinoma with spindle cell metaplasia	Complex epithelial neoplasms
8573	Adenocarcinoma with apocrine metaplasia	Complex epithelial neoplasms
8574	Adenocarcinoma with neuroendocrine differentiation	Complex epithelial neoplasms
8575	Metaplastic carcinoma, NOS	Complex epithelial neoplasms
8576	Hepatoid adenocarcinoma	Complex epithelial neoplasms
8580	Thymoma, malignant, NOS (C37.9)	Thymic epithelial neoplasms
8581	Thymoma, type A, malignant (C37.9)	Thymic epithelial neoplasms
8582	Thymoma, type AB, malignant (C37.9)	Thymic epithelial neoplasms
8583	Thymoma, type B1, malignant (C37.9)	Thymic epithelial neoplasms
8584	Thymoma, type B2, malignant (C37.9)	Thymic epithelial neoplasms
8585	Thymoma, type B3, malignant (C37.9)	Thymic epithelial neoplasms
8586	Thymic carcinoma, NOS (C37.9)	Thymic epithelial neoplasms
8588	Spindle epithelial tumour with thymus-like element	Thymic epithelial neoplasms

8589	Carcinoma showing thymus-like element	Thymic epithelial neoplasms
8600	Thecoma, malignant (C56.9)	Specialized gonadal neoplasms
8620	Granulosa cell tumour, malignant (C56.9)	Specialized gonadal neoplasms
8630	Androblastoma, malignant	Specialized gonadal neoplasms
8631	Sertoli-Leydig cell tumour, poorly differentiated	Specialized gonadal neoplasms
8634	Sertoli-Leydig cell tumour, poorly differentiated, with heterologous elements	Specialized gonadal neoplasms
8640	Sertoli cell carcinoma (C62._)	Specialized gonadal neoplasms
8650	Leydig cell tumour, malignant (C62._)	Specialized gonadal neoplasms
8670	Steroid cell tumour, malignant	Specialized gonadal neoplasms
8680	Paraganglioma, malignant	Paragangliomas and glomus tumours
8693	Extra-adrenal paraganglioma, malignant	Paragangliomas and glomus tumours
8700	Pheochromocytoma, malignant (C74.1)	Paragangliomas and glomus tumours
8710	Glomangiosarcoma	Paragangliomas and glomus tumours
8711	Glomus tumour, malignant	Paragangliomas and glomus tumours
8720	Malignant melanoma, NOS	Nevi and melanomas
8721	Nodular melanoma (C44._)	Nevi and melanomas
8722	Balloon cell melanoma (C44._)	Nevi and melanomas
8723	Malignant melanoma, regressing (C44._)	Nevi and melanomas
8728	Meningeal melanomatosis (C70.9)	Nevi and melanomas
8730	Amelanotic melanoma (C44._)	Nevi and melanomas
8740	Malignant melanoma in junctional nevus (C44._)	Nevi and melanomas
8741	Malignant melanoma in precancerous melanosis (C44._)	Nevi and melanomas
8742	Lentigo maligna melanoma (C44._)	Nevi and melanomas
8743	Superficial spreading melanoma (C44._)	Nevi and melanomas
8744	Acral lentiginous melanoma, malignant (C44._)	Nevi and melanomas
8745	Desmoplastic melanoma malignant (C44._)	Nevi and melanomas
8746	Mucosal lentiginous melanoma	Nevi and melanomas
8761	Malignant melanoma in giant pigmented nevus (C44._)	Nevi and melanomas

8770	Mixed epithelioid and spindle cell melanoma	Nevi and melanomas
8771	Epithelioid cell melanoma	Nevi and melanomas
8772	Spindle cell melanoma, NOS	Nevi and melanomas
8773	Spindle cell melanoma, type A (C69._)	Nevi and melanomas
8774	Spindle cell melanoma, type B (C69._)	Nevi and melanomas
8780	Blue nevus, malignant (C44._)	Nevi and melanomas
8800	Sarcoma, NOS	Nevi and melanomas
8801	Spindle cell sarcoma	Nevi and melanomas
8802	Giant cell sarcoma	Nevi and melanomas
8803	Small cell sarcoma	Nevi and melanomas
8804	Epithelioid sarcoma	Nevi and melanomas
8805	Undifferentiated sarcoma	Nevi and melanomas
8806	Desmoplastic small round cell tumour	Nevi and melanomas
8810	Fibrosarcoma, NOS	Fibromatous neoplasms
8811	Fibromyxosarcoma	Fibromatous neoplasms
8812	Periosteal fibrosarcoma (C40._, C41._)	Fibromatous neoplasms
8813	Fascial fibrosarcoma	Fibromatous neoplasms
8814	Infantile fibrosarcoma	Fibromatous neoplasms
8815	Solitary fibrous tumour, malignant	Fibromatous neoplasms
8830	Malignant fibrous histiocytoma	Fibromatous neoplasms
8832	Dermatofibrosarcoma, NOS (C44._)	Fibromatous neoplasms
8833	Pigmented dermatofibrosarcoma protuberans (C44._)	Fibromatous neoplasms
8840	Myxosarcoma	Myxomatous neoplasms
8850	Liposarcoma, NOS	Lipomatous neoplasms
8851	Liposarcoma, well-differentiated	Lipomatous neoplasms
8852	Myxoid liposarcoma	Lipomatous neoplasms
8853	Round cell liposarcoma	Lipomatous neoplasms
8854	Pleomorphic liposarcoma	Lipomatous neoplasms

8855	Mixed liposarcoma	Lipomatous neoplasms
8857	Fibroblastic liposarcoma	Lipomatous neoplasms
8858	Dedifferentiated liposarcoma	Lipomatous neoplasms
8890	Leiomymosarcoma, NOS	Myomatous neoplasms
8891	Epithelioid leiomyosarcoma	Myomatous neoplasms
8894	Angiomyosarcoma	Myomatous neoplasms
8895	Myosarcoma	Myomatous neoplasms
8896	Myxoid leiomyosarcoma	Myomatous neoplasms
8900	Rhabdomyosarcoma, NOS	Myomatous neoplasms
8901	Pleomorphic rhabdomyosarcoma, adult type	Myomatous neoplasms
8902	Mixed type rhabdomyosarcoma	Myomatous neoplasms
8910	Embryonal rhabdomyosarcoma, NOS	Myomatous neoplasms
8912	Spindle cell rhabdomyosarcoma	Myomatous neoplasms
8920	Alveolar rhabdomyosarcoma	Myomatous neoplasms
8921	Rhabdomyosarcoma with ganglionic differentiation	Myomatous neoplasms
8930	Endometrial stromal sarcoma, NOS (C54.1)	Complex mixed and stromal neoplasms
8931	Endometrial stromal sarcoma, low grade (C54.1)	Complex mixed and stromal neoplasms
8933	Adenosarcoma	Complex mixed and stromal neoplasms
8934	Carcinofibroma	Complex mixed and stromal neoplasms
8935	Stromal sarcoma, NOS	Complex mixed and stromal neoplasms
8936	Gastrointestinal stromal sarcoma	Complex mixed and stromal neoplasms
8940	Mixed tumour, malignant, NOS	Complex mixed and stromal neoplasms
8941	Carcinoma in pleomorphic adenoma (C07._, C08._)	Complex mixed and stromal neoplasms
8950	Mullerian mixed tumour (C54._)	Complex mixed and stromal neoplasms
8951	Mesodermal mixed tumour	Complex mixed and stromal neoplasms
8959	Malignant cystic nephroma (C64.9)	Complex mixed and stromal neoplasms
8960	Nephroblastoma, NOS	Complex mixed and stromal neoplasms
8963	Malignant rhabdoid tumour	Complex mixed and stromal neoplasms

8964	Clear cell sarcoma of kidney (C64.9)	Complex mixed and stromal neoplasms
8970	Hepatoblastoma (C22.0)	Complex mixed and stromal neoplasms
8971	Pancreatoblastoma (C25._)	Complex mixed and stromal neoplasms
8972	Pulmonary blastoma (C34._)	Complex mixed and stromal neoplasms
8973	Pleuropulmonary blastoma	Complex mixed and stromal neoplasms
8980	Carcinosarcoma, NOS	Complex mixed and stromal neoplasms
8981	Carcinosarcoma, embryonal	Complex mixed and stromal neoplasms
8982	Malignant myoepithelioma	Complex mixed and stromal neoplasms
8990	Mesenchymoma, malignant	Complex mixed and stromal neoplasms
8991	Embryonal sarcoma	Complex mixed and stromal neoplasms
9000	Brenner tumour, malignant (C56.9)	Fibroeptihelial neoplasms
9014	Serous adenocarcinofibroma	Fibroeptihelial neoplasms
9015	Mucinous adenocarcinofibroma	Fibroeptihelial neoplasms
9020	Phyllodes tumour, malignant (C50._)	Fibroeptihelial neoplasms
9040	Synovial sarcoma, NOS	Synovial-like neoplasms
9041	Synovial sarcoma, spindle cell	Synovial-like neoplasms
9042	Synovial sarcoma, epithelioid cell	Synovial-like neoplasms
9043	Synovial sarcoma, biphasic	Synovial-like neoplasms
9044	Clear cell sarcoma, NOS	Synovial-like neoplasms
9050	Mesothelioma, malignant	Mesothelial neoplasms
9051	Fibrous mesothelioma, malignant	Mesothelial neoplasms
9052	Epithelioid mesothelioma, malignant	Mesothelial neoplasms
9053	Mesothelioma, biphasic, malignant	Mesothelial neoplasms
9060	Dysgerminoma	Germ cell neoplasms
9061	Seminoma, NOS (C62._)	Germ cell neoplasms
9062	Seminoma, anaplastic (C62._)	Germ cell neoplasms
9063	Spermatocytic seminoma (C62._)	Germ cell neoplasms
9064	Germinoma	Germ cell neoplasms

9065	Germ cell tumour, nonseminomatous (C62._)	Germ cell neoplasms
9070	Embryonal carcinoma, NOS	Germ cell neoplasms
9071	Yolk sac tumour	Germ cell neoplasms
9072	Polyembroma	Germ cell neoplasms
9080	Teratoma, malignant, NOS	Germ cell neoplasms
9081	Teratocarcinoma	Germ cell neoplasms
9082	Malignant teratoma, undifferentiated	Germ cell neoplasms
9083	Malignant teratoma, intermediate	Germ cell neoplasms
9084	Teratoma with malignant transformation	Germ cell neoplasms
9085	Mixed germ cell tumour	Germ cell neoplasms
9090	Struma ovarii, malignant (C56.9)	Germ cell neoplasms
9100	Choriocarcinoma, NOS	Trophoblastic neoplasms
9101	Choriocarcinoma combined with other germ cell elements	Trophoblastic neoplasms
9102	Malignant teratoma, trophoblastic	Trophoblastic neoplasms
9105	Trophoblastic tumour, epithelioid	Trophoblastic neoplasms
9110	Mesonephroma, malignant	Mesonephromas
9120	Hemangiosarcoma	Blood vessel tumours
9124	Kupffer cell sarcoma (C22.0)	Blood vessel tumours
9130	Hemangioendothelioma, malignant	Blood vessel tumours
9133	Epithelioid hemangioendothelioma, malignant	Blood vessel tumours
9140	Kaposi sarcoma	Blood vessel tumours
9150	Hemangiopericytoma, malignant	Blood vessel tumours
9170	Lymphangiosarcoma	Lymphatic vessel tumours
9180	Osteosarcoma, NOS (C40._, C41._)	Osseous and chonrdromatous neoplasms
9181	Chondroblastic osteosarcoma (C40._, C41._)	Osseous and chonrdromatous neoplasms
9182	Fibroblastic osteosarcoma (C40._, C41._)	Osseous and chonrdromatous neoplasms
9183	Telangiectatic osteosarcoma (C40._, C41._)	Osseous and chonrdromatous neoplasms
9184	Osteosarcoma in Paget disease of bone (C40._, C41._)	Osseous and chonrdromatous neoplasms

9185	Small cell osteosarcoma (C40._, C41._)	Osseous and chonrdromatous neoplasms
9186	Central osteosarcoma (C40._, C41._)	Osseous and chonrdromatous neoplasms
9187	Intraosseous well differentiated osteosarcoma (C40._, C41._)	Osseous and chonrdromatous neoplasms
9192	Parosteal osteosarcoma (C40._, C41._)	Osseous and chonrdromatous neoplasms
9193	Periosteal osteosarcoma (C40._, C41._)	Osseous and chonrdromatous neoplasms
9194	High grade surface osteosarcoma (C40._, C41._)	Osseous and chonrdromatous neoplasms
9195	Intracortical osteosarcoma (C40._, C41._)	Osseous and chonrdromatous neoplasms
9220	Chondrosarcoma, NOS (C40._, C41._)	Osseous and chonrdromatous neoplasms
9221	Juxtacortical chondrosarcoma (C40._, C41._)	Osseous and chonrdromatous neoplasms
9230	Chrondroblastoma, malignant (C40._, C41._)	Osseous and chonrdromatous neoplasms
9231	Myxoid chrondrosarcoma	Osseous and chonrdromatous neoplasms
9240	Mesenchymal chondrosarcoma	Osseous and chonrdromatous neoplasms
9242	Clear cell chondrosarcoma (C40._, C41._)	Osseous and chonrdromatous neoplasms
9243	Dedifferentiated chondrosarcoma (C40._, C41._)	Osseous and chonrdromatous neoplasms
9250	Giant cell tumour of bone, malignant (C40._, C41._)	Giant cell tumours
9251	Malignant giant cell tumour of soft parts	Giant cell tumours
9252	Malignant tenosynovial giant cell tumour (C49._)	Giant cell tumours
9260	Ewing sarcoma	Miscellaneous bone tumours
9261	Adamantinoma of long bones (C40._)	Miscellaneous bone tumours
9270	Odontogenic tumour, malignant	Odontogenic tumours
9290	Ameloblastic odontosarcoma	Odontogenic tumours
9310	Ameloblastoma, malignant	Odontogenic tumours
9330	Ameloblastic fibrosarcoma	Odontogenic tumours
9342	Odontogenic carcinosarcoma	Odontogenic tumours
9362	Pineoblastoma (C75.3)	Miscellaneous tumours
9264	Peripheral neuroectodermal tumour	Miscellaneous tumours
9365	Askin tumour	Miscellaneous tumours
9370	Chordoma, NOS	Miscellaneous tumours

9371	Chondroid chordoma	Miscellaneous tumours
9372	Dedifferentiated chordoma	Miscellaneous tumours
9380	Glioma, malignant (C71._)	Gliomas
9381	Gliomatosis cerebri (C71._)	Gliomas
9382	Mixed glioma (C71._)	Gliomas
9390	Choroid plexus carcinoma (C71.5)	Gliomas
9391	Ependymoma, NOS (C71._)	Gliomas
9392	Ependymoma, anaplastic (C71._)	Gliomas
9393	Papillary ependymoma (C71._)	Gliomas
9395	Papillary tumour of the pineal region	Gliomas
9400	Astrocytoma, NOS	Gliomas
9401	Astrocytoma, anaplastic (C71._)	Gliomas
9410	Protoplasmic astrocytoma (C71._)	Gliomas
9411	Gemistocytic astrocytoma (C71._)	Gliomas
9420	Fibrillary astrocytoma (C71._)	Gliomas
9423	Polar spongioblastoma (C71._)	Gliomas
9424	Pleomorphic xanthoastrocytoma (C71._)	Gliomas
9425	Pilomyxoid astrocytoma	Gliomas
9430	Astroblastoma (C71._)	Gliomas
9440	Glioblastoma, NOS (C71._)	Gliomas
9441	Giant cell glioblastoma (C71._)	Gliomas
9442	Gliosarcoma (C71._)	Gliomas
9450	Oligodendroglioma, NOS (C71._)	Gliomas
9451	Oligodendroglioma, anaplastic (C71._)	Gliomas
9460	Oligodendroblastoma (C71._)	Gliomas
9470	Medulloblastoma, NOS (C71.6)	Gliomas
9471	Desmoplastic nodular medulloblastoma (C71.6)	Gliomas
9472	Medullomyoblastoma (C71.6)	Gliomas
9473	Primitive neuroectodermal tumour, NOS	Gliomas

9474	Large cell medulloblastoma (C71.6)	Gliomas
9480	Cerebellar sarcoma, NOS (C71.6)	Gliomas
9493	Dysplastic gangliocytoma of cerebellum	Neuroepitheliomatous neoplasms
9500	Neuroblastoma, NOS	Neuroepitheliomatous neoplasms
9501	Medulloepithelioma, NOS	Neuroepitheliomatous neoplasms
9502	Teratoid medulloepithelioma	Neuroepitheliomatous neoplasms
9503	Neuroepithelioma, NOS	Neuroepitheliomatous neoplasms
9504	Spongioneuroblastoma	Neuroepitheliomatous neoplasms
9505	Ganglioma, anaplastic	Neuroepitheliomatous neoplasms
9508	Atypical teratoid/rhabdoid tumour (C71._)	Neuroepitheliomatous neoplasms
9510	Retinoblastoma, NOS (C69.2)	Neuroepitheliomatous neoplasms
9511	Retinoblastoma, differentiated (C69.2)	Neuroepitheliomatous neoplasms
9512	Retinoblastoma, undifferentiated (C69.2)	Neuroepitheliomatous neoplasms
9513	Retinoblastoma, diffuse (C69.2)	Neuroepitheliomatous neoplasms
9520	Olfactory neurogenic tumour	Neuroepitheliomatous neoplasms
9521	Olfactory neurocytoma (C30.0)	Neuroepitheliomatous neoplasms
9522	Olfactory neuroblastoma (C30.0)	Neuroepitheliomatous neoplasms
9523	Olfactory neuroepithelioma (C30.0)	Neuroepitheliomatous neoplasms
9530	Meningioma, malignant	Meningiomas
9538	Papillary meningioma	Meningiomas
9539	Meningeal sarcomatosis	Meningiomas
9540	Malignant peripheral nerve sheath tumour	Nerve sheath tumours
9560	Neurilemoma, malignant	Nerve sheath tumours
9561	Malignant peripheral nerve sheath tumour with rhabdomyoblastic differentiation	Nerve sheath tumours
9571	Perineurioma, malignant	Nerve sheath tumours
9580	Granular cell tumour, malignant	Granular cell tumours and alveolar soft part sarcomas
9581	Alveolar soft part sarcoma	Granular cell tumours and alveolar soft part sarcomas
9590	Malignant lymphoma, NOS	Malignant lymphomas, NOS or diffuse

9591	Malignant lymphoma, non-Hodgkin, NOS	Malignant lymphomas, NOS or diffuse
9596	Composite Hodgkin and non-Hodgkin lymphoma	Malignant lymphomas, NOS or diffuse
9597	Primary cutaneous follicle centre lymphoma	Malignant lymphomas, NOS or diffuse
9650	Hodgkin lymphoma, NOS	Hodgkin lymphoma
9651	Hodgkin lymphoma, lymphocyte-rich	Hodgkin lymphoma
9652	Hodgkin lymphoma, mixed cellularity, NOS	Hodgkin lymphoma
9653	Hodgkin lymphoma, lymphocyte depletion, NOS	Hodgkin lymphoma
9654	Hodgkin lymphoma, lymphocyte depletion, diffuse fibrosis	Hodgkin lymphoma
9655	Hodgkin lymphoma, lymphocyte depletion, reticular	Hodgkin lymphoma
9659	Hodgkin lymphoma, nodular lymphocyte predominance	Hodgkin lymphoma
9661	Hodgkin granuloma	Hodgkin lymphoma
9662	Hodgkin sarcoma	Hodgkin lymphoma
9663	Hodgkin lymphoma, nodular sclerosis, NOS	Hodgkin lymphoma
9664	Hodgkin lymphoma, nodular sclerosis, cellular phase	Hodgkin lymphoma
9665	Hodgkin lymphoma, nodular sclerosis, grade 1	Hodgkin lymphoma
9667	Hodgkin lymphoma, nodular sclerosis, grade 2	Hodgkin lymphoma
9670	Malignant lymphoma, small B lymphocytic, NOS	Mature B-cell lymphomas
9671	Malignant lymphoma, lymphoplasmacytic	Mature B-cell lymphomas
9673	Mantle cell lymphoma	Mature B-cell lymphomas
9675	Malignant lymphoma, mixed small and large cell, diffuse	Mature B-cell lymphomas
9678	Primary effusion lymphoma	Mature B-cell lymphomas
9679	Mediastinal large B-cell lymphoma (C38.3)	Mature B-cell lymphomas
9680	Malignant lymphoma, large B-cell, diffuse, NOS	Mature B-cell lymphomas
9684	Malignant lymphoma, large B-cell, diffuse, immunoblastic, NOS	Mature B-cell lymphomas
9687	Burkitt lymphoma, NOS	Mature B-cell lymphomas
9688	T-cell/histiocyte rich large B-cell lymphoma	Mature B-cell lymphomas
9689	Splenic marginal zone B-cell lymphoma (C42.2)	Mature B-cell lymphomas
9690	Follicular lymphoma, NOS	Mature B-cell lymphomas

9691	Follicular lymphoma, grade 2	Mature B-cell lymphomas
9695	Follicular lymphoma, grade 1	Mature B-cell lymphomas
9698	Follicular lymphoma, grade 3	Mature B-cell lymphomas
9699	Marginal zone B-cell lymphoma, NOS	Mature B-cell lymphomas
9700	Mycosis fungoides (C44._)	Mature T- and NK-cell lymphomas
9701	Sezary syndrome	Mature T- and NK-cell lymphomas
9702	Mature T-cell lymphoma, NOS	Mature T- and NK-cell lymphomas
9705	Angioimmunoblastic T-cell lymphoma	Mature T- and NK-cell lymphomas
9708	Subcutaneous panniculitis-like T-cell lymphoma	Mature T- and NK-cell lymphomas
9709	Cutaneous T-cell lymphoma, NOS (C44._)	Mature T- and NK-cell lymphomas
9712	Intravascular large B-cell lymphoma (C49.9)	Mature T- and NK-cell lymphomas
9714	Anaplastic large cell lymphoma, T cell and Null cell type	Mature T- and NK-cell lymphomas
9716	Hepatosplenic T-cell lymphoma	Mature T- and NK-cell lymphomas
9717	Intestinal T-cell lymphoma	Mature T- and NK-cell lymphomas
9718	Primary cutaneous CD30+ T-cell lymphoproliferative disorder (C44._)	Mature T- and NK-cell lymphomas
9719	NK/T-cell lymphoma, nasal and nasal-type	Mature T- and NK-cell lymphomas
9724	Systemic EBV positive T-cell lymphoproliferative disease of childhood	Precursor cell lymphoblastic lymphoma
9725	Hydroa vacciniforme-like lymphoma	Precursor cell lymphoblastic lymphoma
9726	Primary cutaneous gamma-delta T-cell lymphoma	Precursor cell lymphoblastic lymphoma
9727	Precursor cell lymphoblastic lymphoma, NOS	Precursor cell lymphoblastic lymphoma
9728	Precursor B-cell lymphoblastic lymphoma	Precursor cell lymphoblastic lymphoma
9729	Precursor T-cell lymphoblastic lymphoma	Precursor cell lymphoblastic lymphoma
9731	Plasmacytoma, NOS	Plasma cell tumours
9732	Multiple myeloma (C42.1)	Plasma cell tumours
9733	Plasma cell leukaemia (C42.1)	Plasma cell tumours
9734	Plasmacytoma, extramedullary	Plasma cell tumours
9735	Plasmablastic lymphoma	Plasma cell tumours
9737	ALK positive large B-cell lymphoma	Plasma cell tumours

9738	Large B-cell lymphoma arising in HHV8-associated multicentric Castleman disease	Plasma cell tumours
9740	Mast cell sarcoma	Mast cell tumours
9741	Malignant mastocytosis	Mast cell tumours
9742	Mast cell leukaemia (C42.1)	Mast cell tumours
9750	Malignant histiocytosis	Neoplasms of histiocytes and accessory lymphoid cells
9751	Langerhans cell histiocytosis, NOS	Neoplasms of histiocytes and accessory lymphoid cells
9755	Histiocytic sarcoma	Neoplasms of histiocytes and accessory lymphoid cells
9756	Langerhans cell sarcoma	Neoplasms of histiocytes and accessory lymphoid cells
9757	Interdigitating dendritic cell sarcoma	Neoplasms of histiocytes and accessory lymphoid cells
9758	Follicular dendritic cell sarcoma	Neoplasms of histiocytes and accessory lymphoid cells
9759	Fibroblastic reticular cell tumour	Neoplasms of histiocytes and accessory lymphoid cells
9760	Immunoproliferative disease, NOS	Immunoproliferative diseases
9761	Waldenstrom macroglobulinemia (C42.0)	Immunoproliferative diseases
9762	Heavy chain disease, NOS	Immunoproliferative diseases
9764	Immunoproliferative small intestinal disease (C17._)	Immunoproliferative diseases
9800	Leukaemia, NOS	Leukaemias, NOS
9801	Acute leukaemia, NOS	Leukaemias, NOS
9805	Acute biphenotypic leukaemia	Leukaemias, NOS
9806	Mixed phenotype acute leukaemia with BCR-ABL1	Leukaemias, NOS
9807	Mixed phenotype acute leukaemia with MLL rearranged	Leukaemias, NOS
9808	Mixed phenotype acute leukaemia, B/myeloid, NOS	Leukaemias, NOS
9809	Mixed phenotype acute leukaemia, T/myeloid, NOS	Leukaemias, NOS
9811	B lymphoblastic leukaemia/lymphoma, NOS	Lymphoid leukaemias
9812	B lymphoblastic leukaemia/lymphoma with BCR-ABL1	Lymphoid leukaemias
9813	B lymphoblastic leukaemia/lymphoma with MLL rearranged	Lymphoid leukaemias
9814	B lymphoblastic leukaemia/lymphoma with TEL-AML1	Lymphoid leukaemias
9815	B lymphoblastic leukaemia/lymphoma with hyperdiploidy	Lymphoid leukaemias
9816	B lymphoblastic leukaemia/lymphoma with hypodiploidy	Lymphoid leukaemias

9817	B lymphoblastic leukaemia/lymphoma with IL3-IGH	Lymphoid leukaemias
9818	B lymphoblastic leukaemia/lymphoma with E2A-PBX1	Lymphoid leukaemias
9820	Lymphoid leukaemia, NOS	Lymphoid leukaemias
9823	B-cell chronic lymphocytic leukaemia/small lymphocytic lymphoma	Lymphoid leukaemias
9826	Burkitt cell leukaemia	Lymphoid leukaemias
9827	Adult T-cell leukaemia/lymphoma (HTLV-1 positive)	Lymphoid leukaemias
9831	T-cell large granular lymphocytic leukaemia	Lymphoid leukaemias
9832	Prolymphocytic leukaemia, NOS	Lymphoid leukaemias
9833	Prolymphocytic leukaemia, B-cell type	Lymphoid leukaemias
9834	Prolymphocytic leukaemia, T-cell type	Lymphoid leukaemias
9835	Precursor cell lymphoblastic leukaemia, NOS	Lymphoid leukaemias
9836	Precursor B-cell lymphoblastic leukaemia	Lymphoid leukaemias
9837	Precursor T-cell lymphoblastic leukaemia	Lymphoid leukaemias
9840	Acute myeloid leukaemia, M6 type	Myeloid leukaemias
9860	Myeloid leukaemia, NOS	Myeloid leukaemias
9861	Acute myeloid leukaemia, NOS	Myeloid leukaemias
9863	Chronic myeloid leukaemia, NOS	Myeloid leukaemias
9865	Acute myeloid leukaemia with DEK-NUP214	Myeloid leukaemias
9866	Acute promyelocytic leukaemia	Myeloid leukaemias
9867	Acute myelomonocytic leukaemia	Myeloid leukaemias
9869	Acute myeloid leukaemia	Myeloid leukaemias
9870	Acute basophilic leukaemia	Myeloid leukaemias
9871	Acute myeloid leukaemia with abnormal marrow eosinophils	Myeloid leukaemias
9872	Acute myeloid leukaemia, minimal differentiation	Myeloid leukaemias
9873	Acute myeloid leukaemia without maturation	Myeloid leukaemias
9874	Acute myeloid leukaemia with maturation	Myeloid leukaemias
9875	Chronic myelogenous leukaemia, BCR/ABL positive	Myeloid leukaemias
9876	Atypical chronic myeloid leukaemia, BCR/ABL negative	Myeloid leukaemias
9891	Acute monocytic leukaemia	Myeloid leukaemias

9895	Acute myeloid leukaemia with myelodysplasia-related changes	Myeloid leukaemias
9896	Acute myeloid leukaemia t(8;21)(q22;q22)	Myeloid leukaemias
9897	Acute myeloid leukaemia, 11q23 abnormalities	Myeloid leukaemias
9898	Myeloid leukaemia associated with Down Syndrome	Myeloid leukaemias
9910	Acute megakaryoblastic leukaemia	Myeloid leukaemias
9911	Acute myeloid leukaemia (megakaryoblastic) t(1;22) (p13;q13) RBM15-MKL1	Myeloid leukaemias
9920	Therapy related myeloid neoplasm	Myeloid leukaemias
9930	Myeloid sarcoma	Myeloid leukaemias
9931	Acute panmyelosis with myelofibrosis (C42.1)	Myeloid leukaemias
9940	Hairy cell leukaemia (C42.1)	Myeloid leukaemias
9945	Chronic myelonomocytic leukaemia, NOS	Other leukaemias
9946	Juvenile myelomonocytic leukaemia	Other leukaemias
9948	Aggressive NK-Cell leukaemia	Other leuakemias
9950	Polycythemia vera	Chronic myeloproliferative disorders
9960	Myeloproliferative neoplasm, NOS	Chronic myeloproliferative disorders
9961	Primary myelofibrosis	Chronic myeloproliferative disorders
9962	Essential thrombocythemia	Chronic myeloproliferative disorders
9963	Chronic neutrophilic leukaemia	Chronic myeloproliferative disorders
9964	Chronic eosinophilic leukaemia, NOS	Chronic myeloproliferative disorders
9965	Myeloid and lymphoid neoplasms with PDGFRA rearrangement	Chronic myeloproliferative disorders
9966	Myeloid neoplasms with PDGFRB rearrangement	Chronic myeloproliferative disorders
9967	Myeloid and lymphoid neoplasms with FGFR1 abnormalities	Chronic myeloproliferative disorders
9971	Polymorphic post-transplant lymphoproliferative disorder	Other hematologic disorders
9975	Myeloproliferative neoplasm, unclassifiable	Other hematologic disorders
9980	Refractory anaemia	Myelodysplastic syndromes
9982	Refractory anaemia with sideroblasts	Myelodysplastic syndromes
9983	Refractory anaemia with excess blasts	Myelodysplastic syndromes
9984	Refractory anaemia with excess blasts in transformation	Myelodysplastic syndromes

9985	Refractory cytopenia with multilineage dysplasia	Myelodysplastic syndromes
9986	Myelodysplastic syndrome with 5q deletion (5q-) syndrome	Myelodysplastic syndromes
9987	Therapy-related myelodysplastic syndrome, NOS	Myelodysplastic syndromes
9989	Myelodysplastic syndrome, NOS	Myelodysplastic syndromes
9991	Refractory neutropenia	Myelodysplastic syndromes
9992	Refractory thrombocytopenia	Myelodysplastic syndromes