

Data Explained

Cancer Network Information System (CNIS)

Authors: Dr Laura Baker, Dr Arron Lacey, Professor Ashley Akbari

Date: February 2025

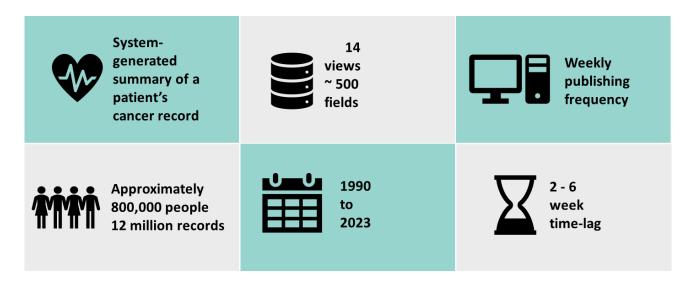
This Data Explained summarises the Cancer Network Information System (CaNISC – known as CNIS in the SAIL Databank) data. This output is intended to help guide users interested in this data for the first time, and future use of the data within the SAIL Databank towards research delivery and to provide feedback into future data source development and documentation.

The data discussed in this Data Explained was made securely available through an approved SAIL project via the independent Information Governance Review Panel (IGRP) – project 1598. The data used in this approved SAIL project comes from Digital Health and Care Wales (DHCW) and was accessed through the Secure Anonymised Information Linkage (SAIL) Databank. The data was not originally collected for research and it is expected that there are gaps and inconsistencies in its recording, a number of which are detailed in the following. The work presented in this Data Explained is correct at the point of publication. Views expressed in this Data Explained are those of the researchers and not necessarily those of ADR Wales partner organisations.

*

Overview

The Cancer Network Information System Cymru (CaNISC) is a data repository for all aspects of cancer diagnosis and treatment, which is accessible via the SAIL Databank. The infographic below summarises the data sources' features at the time of this publication.



Introduction

Cancer is a complex and significant public health issue, being a leading cause of death globally (Siegel et al., 2023; WHO, 2020). The increasing incidence of cancer, combined with an ageing population, places greater demand on healthcare services and highlights the need for effective detection, treatment, and intervention strategies.

The Cancer Network Information System Cymru (CaNISC, known as the CNIS data in the SAIL Databank) is a data repository for all aspects of cancer diagnosis and treatment. This system is the main source of data when evaluating patients' cancer diagnoses, care, and the wait times they encounter. CNIS informs cancer registration, providing a patient's cancer care data from diagnosis through to progression (Morgan, 2023).

All Health Boards and Trusts, hospices, and other organisations involved in delivering cancer care in Wales use CNIS. It enables multiple organisations to record and share information about a patient's diagnosis, treatment, and follow-up care. CNIS collects data in a live setting and, as such, requires significant post-processing to understand and interpret for research purposes. It can in turn be cleaned, curated and derived into a data source that can then be used for research and/or as a research-ready data asset (RRDA).

CNIS has notable limitations. For example, it is only accessible to approximately 2,000 healthcare professionals (for example, cancer services, specialist palliative care services, screening and registry teams). It operates as a standalone application, meaning clinical information is only visible via CNIS access (NHS Wales, 2021). Consequently, when patients receive care outside of their cancer centres, the associated clinical information is not accessible to other healthcare providers. To address these issues, the Cancer Informatics Programme (CIP) is currently being rolled out in

Wales to replace CNIS across all Health Boards and Trusts (Morgan, 2023). The CIP aims to enhance existing interfaces in the NHS (including the Welsh Patient Administration System (WPAS) and Welsh Clinical Portal (WCP)) to improve the functionality of CNIS and provide all clinicians access to a patient's full healthcare record.

How is the data collected?

Data in CNIS is collected from primary care and secondary care via accident and emergency (A&E, also known as emergency department (ED)), hospital admissions (also known as in-patients) and outpatients, and can be used to follow all pathways associated with cancer (HDR UK, 2024). CNIS holds data that "includes multidisciplinary team diagnosis, proposed treatments, and a system-generated summary of the patient's cancer record" (HDR UK, 2024).

However, there is a notable depletion in documentation regarding the data flow from collection to deposition in CNIS in the public domain. As such, any processes between deposition within primary care, etc. to CNIS are currently unknown at this time.

Data Linkage

Within the SAIL Databank, the CNIS data source consists of 14 data views (**Figure 1**), with linkage between the views facilitated by the Anonymised Linking Field (ALF).



_

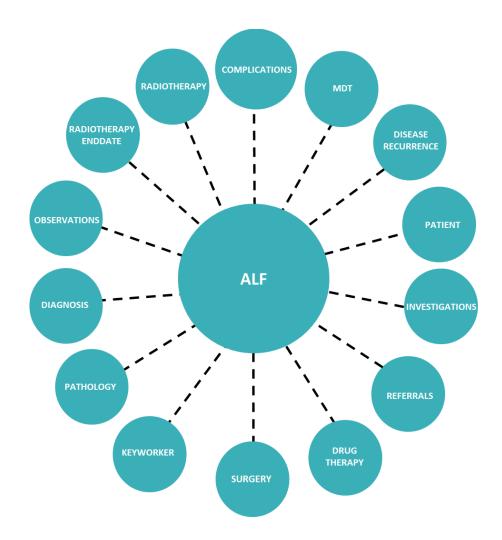


Figure 1. Overview of data views and common linkage fields in the CNIS data source.

Accessing CNIS through the SAIL Databank allows cancer diagnostic and outcomes data to be linked with other data sources, including routinely collected electronic health record (EHR), administrative, survey, registry and other data sources. A full list of data sources available via the Databank is provided here:

https://healthdatagateway.org/en/search?type=datasets&sampleAvailability=&publisherName=SAIL.

This data linkage supports the development of more specific and targeted research questions. For example, it enables studies on how cancer diagnosis might influence the onset of other comorbidities, as well as many other ways in which linkage could be used to evaluate trends at the population level.

What can the data be used for?

CNIS structure and content

CNIS contains detailed individual-level records that document each person's journey through the healthcare system following a cancer diagnosis. Each record is linked via an encrypted NHS number (ALF within SAIL). There is a one-to-many relationship between the different views, as multiple records can exist for each treatment, surgery, pathology test, or other healthcare interactions.

The specific criteria for data collection and deposition are not currently available to the public, nor is the rationale behind the organisation of CNIS data into the views shown in **Figure 1** fully explained. However, it appears that an individual's cancer pathway within CNIS can be traced across separate categories (**Figure 2**). For example, demographic details are contained within the **CNIS.patient** view (**Figure 2**), whereas treatment details are separated into treatment specific views (e.g., **CNIS.surgery** and **CNIS.drug_Therapy**).

Across the 14 views, record counts range from approximately 36,000 to 1.6 million (**Table A1**), representing a cohort of 780,000 individuals, with data available in SAIL up to 2024 (**Table A1**). Depending on its purpose, the cohort size varies by view. For instance, the largest cohorts are found in the Patient, Diagnosis, and Referral views, as these represent stages all individuals typically experience during their cancer journey. Conversely, smaller cohorts are found in more specialised views, such as Treatment, Surgery, and Drug Therapy, as not all individuals undergo chemotherapy or surgery. The variation in cohort sizes across views suggests that individuals may enter the CNIS data source at multiple points (**Figure 2**).

There is also notable variation in the time coverage between views (**Table A1**). It is highly unlikely that data from the early 1900s are correct, given the CaNISC system was not active at that time (albeit data may have been added retrospectively). As such, a data-cleaning step is recommended to remove outlier dates from the analysis.



1

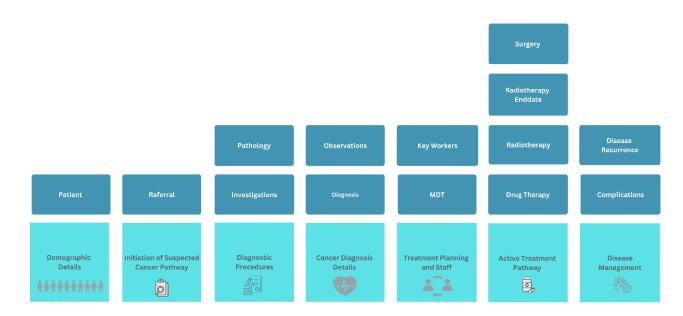


Figure 2. The anticipated CNIS data pathway for individuals suspected of cancer.

The pathway an individual follows within the CNIS data views when cancer is suspected (dark blue) and the associated features of the data (light blue).

Exploring the data pathway

An individual's journey typically starts with their information being recorded in the *CNIS.Patient* view (Figure 2). This view captures demographic details, such as the individual's address and date of birth. However, the methods for collecting and verifying this information are currently unclear.

Pathology testing

When an illness is suspected, many individuals undergo pathology testing, which is documented in the *CNIS.pathology* view. This view tracks the encrypted NHS number (ALF) across various pathology tests, such as histology core investigations (e.g., biopsy/excision), along with corresponding clinical codes (from an unspecified coding system), the dates samples were received, and the dates results were issued. In addition, specific pathology results are available, including details such as the suspected cancer type and ICD-10 code (Terminology and Classifications Delivery Service, 2024). Further testing data, including endoscopy-related procedures, can be found in the *CNIS.Investigations* view, although this view has limited coverage.

Referral process

Testing can take place at various levels, from primary to secondary care, following an initial referral. *CNIS.referral* view allows for the analysis of referral types (e.g., from a general practitioner (GP) or secondary care consultants), referral priority status, and subsequent appointment information, including the first offered appointment and the first appointment where the individual was seen (**Figure 2**). This view also records delays, communication routes for referrals, and the date the individual was informed of their diagnosis.

Cancer diagnosis

When an individual is diagnosed with cancer, details are recorded in the *CNIS_Diagnosis* view. This includes:

- Information on the diagnosis presentation
- Laterality
- Grade
- Stage
- Stage type
- Basis of diagnosis
- Date of diagnosis

Observations made at the time of diagnosis, such as the individual's weight and height, can be found in the *CNIS.observations* view (Figure 2).

Multidisciplinary team (MDT) meetings

Following diagnosis, treatment plans are typically discussed by a multidisciplinary team (MDT). The *CNIS.MDT* view contains details about MDT meetings, including planning dates, treatment intentions, treatment plans (e.g., chemotherapy, surgery), and comorbidity scores (**Figure 2**). However, there is uncertainty about the comorbidity scoring system used, as the data dictionary does not specify whether this relates to the <u>Charlson Comorbidity Index</u> (Roffman et al., 2016) or another scoring system or index. The *MDT discussion* detail field is empty and should not be relied upon for research.

Treatment plans

Treatment plans can be further examined in three specific views:

CNIS.radiotherapy/radiotherapy_enddate:

Includes details on the type of radiotherapy prescribed, the prescribed versus actual doses administered, the start date, both prescribed and administered fractions, treatment intention, and the treatment site (Figure 2).

CNIS.drug_therapy:

Documents the intention of the treatment prescribed, the type of drug therapy administered and the start and end dates of treatment (Figure 2).

CNIS.surgery:

Contains information about the intention of surgery being undertaken, the urgency of the procedure (e.g., elective or emergency), the respective OPCS Classification of Interventions and Procedures (OPCS-4S) code of the procedure being undertaken (NHS Digital, 2024), the grade of cancer at the point of surgery, the admission, procedure and discharge dates, the surgical access method undertaken (e.g., laparoscopic or open), the surgical side, an encrypted surgeon code, and whether post-operative death occurred (yes/no field).

Complications and recurrence

Any complications during the cancer treatment pathway, such as the cause, date, and relevant comments, are recorded in the *CNIS.complication's* view. In addition, the recurrence of the disease is tracked in the *CNIS.Disease_recurrence* view, where it is possible to identify if and when the disease has reoccurred, how it was detected, whether it was identified post-treatment and whether it was the individual's first relapse (yes/no) field (Figure 2).

Variable Insight

The following fields are consistent across all 14 views of CNIS:

- alf_pe a double encrypted unique individual identifier.
- pat_id_pe encrypted ID of patient.
- Diaginstance diagnosis incidence
- diag_vrs diagnosisversion
- mix_dt most accurate diagnosis date from a mixture of sources
- site_cd site code
- prov_name provider name
- prov_cd provider code
- epi_type episode type
- epi_type_cd episode type code
- first_inserted_dtlast_updated_dt date last updated

At the time of writing, data definitions for most of these fields (with the exception of alf_pe), are either blank or vague, offering little context regarding their relevance. For example, pat_id_pe in the data dictionary is only described as an encrypted ID of a patient, without providing further details regarding the encryption method and how a patient ID is assigned within the original system. The field mix_dt appears to correspond with the diagnosis date, but the data dictionary does not provide additional details about this field.

Further details and meta-data on the CNIS data can be found here: https://healthdatagateway.org/en/dataset/358

Data Use Examples

The 14 views within CNIS can be individually analysed or combined with others to address specific research questions. Currently, there is no published research explicitly using CNIS as a data source to examine cancer trends.

CNIS offers numerous opportunities to explore trends in cancer diagnosis, treatment, progression, and outcomes. To illustrate the potential of this data source and for the purpose of this data explained, several examples (predominately using the diagnosis table as an exemplar) were selected to illustrate the potential of this data source.

Example 1. Demographic trends

Similar to registry data, CNIS allows for the examination of the incidence of cancer diagnoses over time. This can provide additional information and insights to users of the data interested in looking at longitudinal data and changes in information over time by comparing the data with official registered diagnoses in WCISU and identifying discrepancies between official statistics and what is visible to oncologists on specific databases (such as CNIS).

By joining the *site_cd* field in *CNIS.diagnosis* with ICD-10 codes, it is possible to identify cancer diagnoses in the data. Following this, the data was streamlined to focus on the four most common cancers in Wales identified by the WCISU (<u>WCISU, 2022</u>):

Colorectal: C18 - C20Lung: C33 - C34Breast: C50

• Prostate: C61

Unique combinations of each individual (*alf_pe*), the *location* (*site_cd*), and diagnosis date (*diagdatediag*) were identified in the *CNIS.diagnosis* view. The *site_cd* codes were filtered to focus on breast, prostate, colorectal, and lung cancers. Counts of records per year for each cancer type were compiled and filtered to exclude any counts below 10 to avoid disclosure. The analysis was restricted to the period from 1990 to 2022 to ensure the best coverage (**Figure 3**).



_

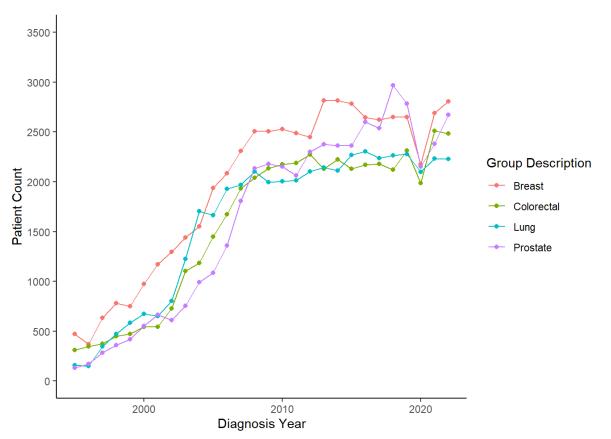


Figure 3. A longitudinal analysis of prevalent cancers in Wales. The CNIS data was utilised to identify distinct individuals diagnosed with Breast, Colorectal, Lung, and Prostate cancer annually.

This approach can be refined further by applying more specific criteria for classifying a cancer case, such as aligning with WCISU cancer registration standards. Additionally, further cleaning rules, such as assessing the quality of ALF matching, can be implemented depending on the study's scope and objectives.

Example 2. Trends outside the scope of cancer registry data

As CNIS serves as a data source that feeds into registry data, CNIS includes not only primary data collected for cancer registration, but also additional parameters not included in the cancer registry's minimum dataset (<u>Public Health Wales NHS Trust, 2024</u>). For example, CNIS provides data on grading at different stages (e.g., at diagnosis and surgery), details of surgical interventions, treatments such as drug therapies and radiotherapy, and related complications.

This makes CNIS valuable as a standalone data source for identifying trends in areas not covered by the registry, offering more detailed and timely insights that may support the analysis of more specific research questions. For instance, tracking changes in cancer grading over time, analysing treatment frequencies and outcome trends, and monitoring the incidence of secondary cancers. However, unlike registry data, which undergoes extensive cleaning and completion processes (which generates a time lag of approximately 18 months), CNIS data may be less reliable due to a less rigorous data cleaning and validation process, as well as a lack of thorough documentation detailing the data's journey through CaNISC systems.

Example 2.1. The methods used to make a diagnosis

A simple count of individuals by diagnosis method (diagpresentation) was conducted using the CNIS.diagnosis view. This revealed data containing approximately 410,000 new diagnoses episodes, 69,500 clinically detected, 25,000 screen detected, 10,000 cases of recurrent disease, 10,000 cases of long-standing disease, and 38,200 others, along with around 516,000 null values.

Example 2.2. Residual disease post-surgery

The *CNIS.surgery* view includes a field that provides information on the occurrence of residual disease remaining after surgery. The frequency of different residual disease sizes is illustrated in Figure 4.



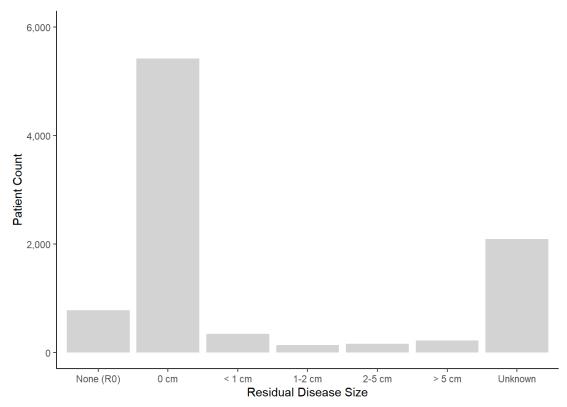


Figure 4. The frequency of residual disease size post-surgery in individuals with cancer

Frequency per presentation were derived from the aggregation of unique encrypted patient numbers (ALF) as presented in the *CNIS.diagnosis* view. Counts are grouped according to the size categories of residual disease, as detailed in the *CNIS.surgery* view, only including individuals who had a recorded value (excluding 412,430 null values). Please note all available data has been presented including potential outliers to showcase the data available.

Example 2.3. Complications

Other views, such as *CNIS.complications*, offer detailed insights into complications arising from cancer treatments. For example, this view highlights the frequency of different complications. This view has data that shows approximately 4,300 complications occurring post-surgery, 700 after radiotherapy, and 300 following chemotherapy in the entire dataset. It is important to note that this field contains about 110,800 null values, which may reflect individuals within the CNIS data who did not experience any complications.

These examples were selected to illustrate the potential uses of the CNIS data. However, CNIS is extensive, and there are numerous other fields that may enhance study design, depending on the research question. Different views and fields may offer additional insights or highlight aspects of cancer diagnosis and treatment that could be valuable for specific analysis.

Example 3. Cancer grading and staging information held in CNIS

Cancer staging classifies a tumour from stage *I* to *IV* based on its size and spread. Cancer grading assesses how abnormal the cells are, also ranging from 1 to 4, with higher grades indicating a greater likelihood of cancer spreading. CNIS provides data on both cancer staging and grading, allowing for detailed analysis. Not all cancers get staging information, e.g., blood, brain and small cell cancers.

Staging

In the *CNIS.diagnosis* view, staging data shows approximately:

- 12,000 individuals at Stage 0.
- 70,300 at Stage I.
- 91,500 at Stage II.
- 74,300 at Stage III.
- 73,500 at Stage IV.
- 54,000 with unknown stages.
- 1.3 million null values.
- 100 cases of Occult Carcinoma.

Grading

Grading information is available in the *CNIS.diagnosis*, *CNIS.pathology*, and *CNIS.surgery* views. At diagnosis, the CNIS data contains approximately:

- 1,000 borderline cases.
- 39,000 Grade I cases.
- 96,000 Grade 2 cases.
- 60,000 Grade 3 cases.
- 3,000 Grade 4 cases.
- 120,000 unknowns.

Additionally, changes in the incidence of cancer stages and grades at diagnosis over time can be tracked (**Figure 5**). To ensure accuracy, distinct records were identified based on unique combinations of the patient identifier (*alf_pe*), diagnosis date (*datediagdate*), and staging information (*diagstagegrouping*) from the **CNIS.diagnosis** view for staging (**A**) or grading information (*diaggrade*) for grading (B). This approach accounts for individuals with multiple staging or grading entries.

Data cleaning was necessary to standardise variations in the staging and grading fields, ensuring consistent grouping. For instance, different representations of the same stage (e.g., "Stage III," "stage III," and "Stage 3") and grade (e.g., "Grade IV," "grade IV," and "grade iv") were transformed into a uniform format. Individuals were then counted per year based on their stage group.

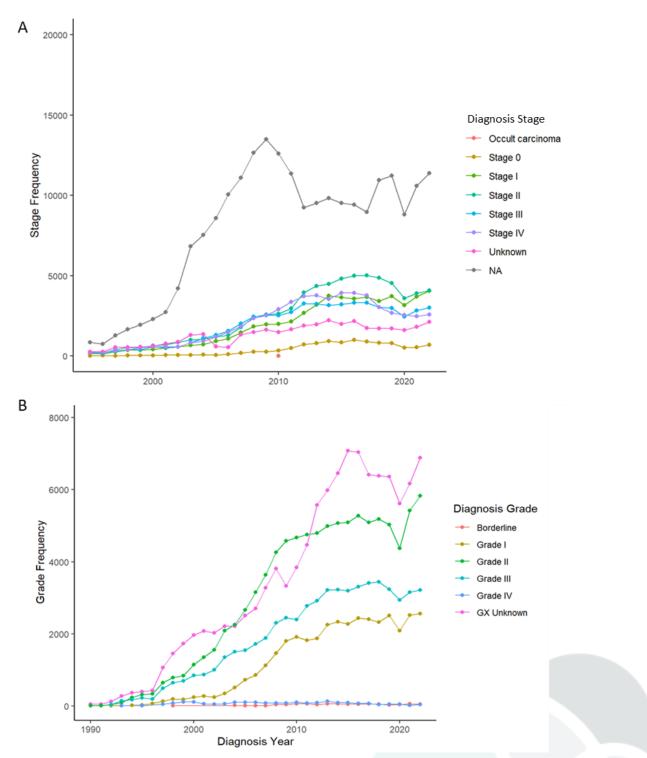


Figure 5. A longitudinal analysis of cancer grading and staging groups in Wales.

Annual counts of distinct ALFs by cancer grading/staging group.

A. Staging groups were identified in the CNIS.diagnosis data, and distinct ALFs were counted per year.

B. Grading groups were identified in the *CNIS.diagnosis* data, and distinct ALFs were counted per year. Counts of 10 or fewer were excluded from the analysis.

Data limitations encountered and suggested improvements

The restrictions regarding selective clinical access to CaNISC is a documented limitation, which has ultimately resulted in its decommission within NHS Wales. In addition, a significant limitation of CNIS is the lack of publicly available documentation. For example, several fields are not research-ready due to insufficient descriptions or the absence of lookup tables. In addition, the existing data dictionary has many fields marked as blank (<a href="https://horsity.com

Moreover, the metadata does not include information on the data processing journey, making it difficult to understand how the data was handled within the CNIS repository, whether any cleaning was performed, and the reasoning behind the data's distribution across different views.

Furthermore, when analysing the data, there is inconsistency in individual-level data across views and varying timeframes within those views. While this may reflect differences in individual cancer journeys, more detailed information about these parameters would be beneficial for researchers considering the use of CNIS in their studies. It is therefore recommended that data owners make this information publicly available.

To enhance this data source, it would be beneficial to include distinct *treatment_ids* to differentiate between various treatment cycles (such as in drug therapy). This would facilitate the analysis of treatment efficacy over time. Additionally, incorporating a field that indicates whether the cancer has metastasised would be valuable for tracking cancer progression and understanding the types of cancer more comprehensively.

Summary

In summary, the CNIS data source provides a valuable resource for exploring additional parameters in population-level cancer research. Its breadth offers many useful fields that, when combined with WCISU, can help investigate various aspects of cancer detection, diagnosis, treatment, and prognosis. However, caution is necessary when assessing field coverage due to inconsistencies and gaps in metadata, which can make evaluating the data's utility challenging. Furthermore, the CaNISC system is in the process of transitioning to a new system, the WCP, as part of the CIP. This transition will likely limit the availability of CaNISC data to a specific timeframe, and any new data generated from this change will need to be analysed from updated perspectives.

Despite these limitations, CNIS holds over 30 years of individual level detailed cancer records, which can be used to inform retrospective cancer studies. In addition, when used alongside cancer registry data, CNIS can support more specific and focused research questions.

References

- HDR UK. (2024). Cancer Network Information System (CNIS / CANISC). https://web.www.healthdatagateway.org/dataset/da91f0eb-1761-485b-8b7f-9121b6a1d004
- Morgan, E. (2023, January 30). Written Statement: Update on progress of the Cancer Informatics Programme. GOV.WALES. https://www.gov.wales/written-statement-update-progress-cancer-informatics-programme
- NHS Digital. (2024). OPCS Classification of Interventions and Procedures [Dita]. Data Dictionary NHS. https://www.datadictionary.nhs.uk/supporting_information/opcs_classification_of_interventions and procedures.html
- NHS Wales. (2021, March 26). What is Canisc? NHS Wales Executive. https://executive.nhs.wales/functions/networks-and-planning/cancer/clinical-hub/canisc/what-is-canisc/
- Public Health Wales NHS Trust. (2024). WCISU Cancer Reporting Tool Technical Guide. https://phw.nhs.wales/services-and-teams/welsh-cancer-intelligence-and-surveillance-unit-wcisu/cancer-reporting-tool-official-statistics/
- Roffman, C. E., Buchanan, J., & Allison, G. T. (2016). Charlson Comorbidities Index. *Journal of Physiotherapy*, 62(3), 171. https://doi.org/10.1016/j.jphys.2016.05.008
- Siegel, R. L., Miller, K. D., Wagle, N. S., & Jemal, A. (2023). Cancer statistics, 2023. CA: A Cancer Journal for Clinicians, 73(1), 17–48. https://doi.org/10.3322/caac.21763
- Terminology and Classifications Delivery Service. (2024). *ICD-10 2024 5th Ed NCCS* (pp. 1–304). https://classbrowser.nhs.uk/ref_books/ICD-10_2024_5th_Ed_NCCS.pdf
- WCISU. (2022). Trends in cancer incidence and stage at diagnosis in Wales up to 2019. Public Health Wales. https://phw.nhs.wales/news/trends-in-cancer-incidence-and-stage-at-diagnosis-in-wales-up-to-2019/
- WHO. (2020). *The top 10 causes of death*. World Health Organisation. https://www.who.int/news-room/fact-sheets/detail/the-top-10-causes-of-death



Glossary

Abbreviation	Name		
ADRUK	Administrative Data Research UK		
ALF	Anonymised Linkage Field		
A&E	Accidents and Emergencies		
CIP	Cancer Informatics Programme		
CNIS	CaNISC - Cancer Network Information System		
CReSt	The Cancer Research Strategy for Wales		
DHCW	Digital Health and Care Wales		
ED	Emergency Department		
ESRC	Economic and Social Research Council		
GP	General Practitioner		
HDR UK	Health Data Research United Kingdom		
IDC-10	International Classification of Diseases Version 10		
IGRP	Information Governance Review Panel		
MDT	MultidisciplinaryTeam		
NHS	National Health Service		
OPCS	Classification of Interventions and Procedures (OPCS-4S)		
ONS	Office for National Statistics		
PHE	Public Health England		

PHW	Public Health Wales
RRDA	Research Ready Data Asset
SAIL	Secure Anonymised Information Linkage (Databank)
TRE	Trusted Research Environment
WCP	Welsh Clinical Portal
WCRC	Wales Cancer Research Centre
WCSU	WCISU - Welsh Cancer Intelligence and Surveillance Unit
WPAS	Welsh Patient Administration System



Disclaimer

This work was produced using administrative data accessed through The SAIL Databank. The use of the data in this work does not imply the endorsement of SAIL or data owners in relation to the interpretation or analysis. This work uses research data sources which may not exactly reproduce National Statistics aggregates. National Statistics follow consistent statistical conventions over time and cannot be compared to Data First linked data sources.

Acknowledgements

This work is supported by ADR Wales, part of ADR UK (Administrative Data Research UK), a partnership transforming access to public sector data in the UK to inform policy decisions that enhance people's lives. ADR UK is funded by the Economic and Social Research Council (ESRC), part of UK Research and Innovation ES/W012227/1]. The research was facilitated by the SAIL Databank, with funding from Health and Care Research Wales via the Wales Cancer Research Centre's CReSt catalytic funding (researcher salary and data access costs) and funding from the Roche Bioresource Data Accelerator programme (data access costs).

We gratefully acknowledge the contributions and support of Professor Sunil Dolwani, Professor Mererid Evans, Dr Mark Davies, Liz Merrifield, Dr Peter Giles and Jenni Macdougall.

Contact

Name: Laura Baker

Email: l.m.baker@swansea.ac.uk



Appendix

 $\textbf{Table A1.} Total \ number \ of \ records, distinct \ individual \ counts \ (\textit{alf} \ and \ \textit{patient_id}) \ per \ CNIS \ view.$

	Record	Distinct	Distinct Individuals
View Name	Count	Individuals (ALF)	(Patient_ID)
Complications	160,816	115,903	115,904
Diagnosis	1,621,832	777,731	780,210
Disease Recurrence	36,300	18,580	18,580
Drug Therapy	845,215	113,529	113,535
Investigations	1,595,034	437,342	437,353
Keyworker	1,528,575	777,731	780,210
MDT	1,203,637	362,585	362,599
Observations	518,443	166,510	166,517
Pathology	243,938	160,971	160,976
Patient	1,531,805	777,731	780,210
Radiotherapy	222,640	142,696	143,927
Radiotherapy Enddate	206,189	142,696	143,927
Referral	1,535,365	777,731	780,210
Surgery	434,379	183,742	183,752



