Data Explained

The Welsh Results Reports Service (WRRS) Data

Wales Population-Scale Pathology Data, A National Data Asset

Authors: Gareth Davies, Ashley Akbari, Hoda Abbasizanjani, Stuart Bedston, Victoria Best, Jane Lyons, Fatemeh Torabi, Ronan Lyons
on behalf of the Swansea Multi-Morbidity and COVID Con-Cov teams.

Date: April 2022

Contents

Introduction

The Welsh Results Reporting Service (WRRS) allows health care professionals (HCPs) across Wales to access, enter and view laboratory results for pathology requests and any other associated results across all health boards in Wales, from both primary and secondary care, regardless of where they were requested, tested or provided back to patients in Wales. WRRS aims to save time, reduce test duplication and improve patient safety. The results are viewed through the Welsh Clinical Portal [1].

The Welsh Clinical Portal (WCP) is a digital patient record across NHS Wales, which is available to all HCPs with appropriate permissions in relevant organisations. The WCP makes available patient information from several sources with a single log-on. The WCP gives clinicians pathology results (e.g. blood tests) for patients wherever they had their test taken, meaning patients can utilise mobile units or local centres rather than having to travel far [1].

How the pathology data are collected

WRRS is taken (indirectly) from the Welsh Laboratory Information Management System (WLIMS), which covers all of Wales. WLIMS is a clinical IT system used by pathology staff across Wales for storing, recording and exchanging pathology laboratory results. WLIMS also links to the machines which conduct the tests and analyse the samples. The system is linked to analysers used to produce the majority of tests within laboratories.

Data location:

The WLIMS is centrally hosted but provides the service to every Health Board laboratory and other national specialist support services, including (as of January 2021):

- Haematology.
- Biochemistry.
- Immunology.
- Cervical Cytology.
- Infection Control Services.
- Mortuary Services.
- Blood Transfusion (to be implemented into WLIMS).
- Haemonetics (blood tracking) (to be implemented into WLIMS).

Improvements to services:

The national system of WLIMS has improved services by introducing a standard approach to testing. It allows HCPs to see all previous tests conducted for a patient, and request new tests, no matter where they are in Wales.

The WLIMS links to the WCP and includes functions to support [2]:

- Improved clinical governance.
- Flexibility, portability and adaptability of service provision.
- Improved demand management and forward planning.
- An improved patient experience.
- Service standardisation.
- Provision of information to the Welsh general practice record.

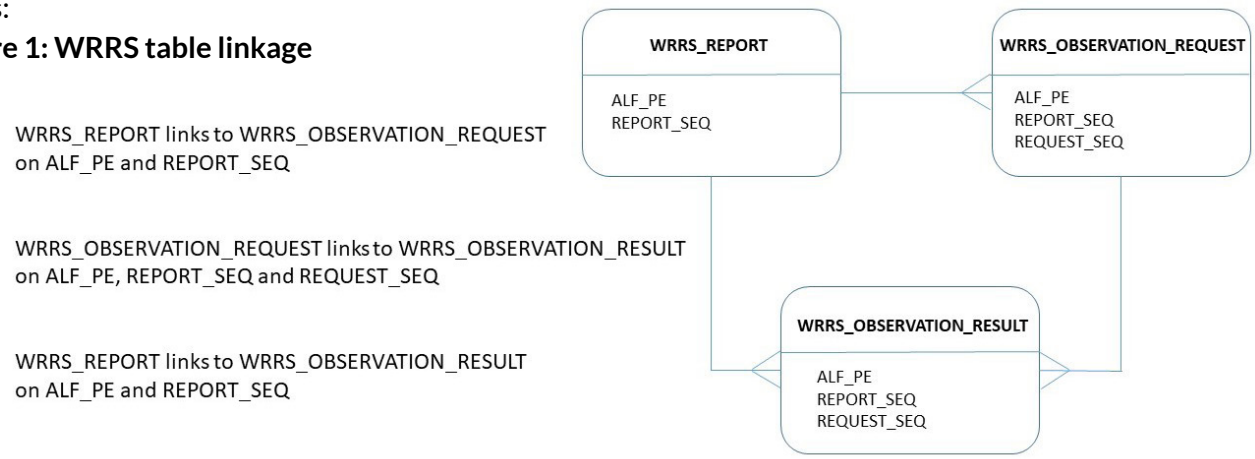
WRRS data are provisioned into the Secure Anonymised Information Linkage (SAIL) Databank by Digital

Health and Care Wales (DHCW). DHCW provide SAIL's trusted third party (TTP) encryption service, ensuring all resultant data acquired into SAIL are anonymised. The WRRS data is a subset of WLIMS and currently contains the pathology results of WLIMS.

What the pathology data can be used for

WRRS currently contains pathology results from WLIMS, but there is the potential to expand the scope of the data asset further in future years based on research need and data quality. The WRRS data in SAIL are divided into three separate tables/views, and can be linked together by SAIL users using the following linkage fields:

Figure 1: WRRS table linkage



The currently available fields in the WRRS data within SAIL are listed and described under the "WRRS metadata" heading in the supplementary material. They are also available on the SAIL meta-data catalogue and HDR innovation gateway (insert URL)

Defining research areas of interest in WRRS

The WRRS data does not currently have an associated data dictionary to describe and explain the contents to users. Therefore, the Population Data Science group at Swansea University aims to establish a data dictionary that would collate meta-data and generate a streamlined process to create harmonised code lists over time. Through our use of the WRRS data in the delivery of research, we hope to share our learning with other SAIL users in order to enable others to have a more efficient experience in accessing and extracting relevant information from the WRRS through the reuse of reproducible code lists and documentation. We also invite other users of SAIL to collaborate with us in contribution to the data dictionary with new code lists, concepts and learning. Finally, we hope to apply our understanding of the WRRS to create a research ready data asset (RRDA) in the future, which will create a more streamlined, clean research ready version of the data with the code lists and dictionary pre-applied.

The data dictionary is an iterative development, with a streamlined process having been established and adopted by the group, and is actively being implemented on a per research project basis for each code list/ concept/condition/outcome definition of interest at a time. As each research area of interest is considered, the list of details extracted from the WRRS, such as biomarkers, test names, test results, clinical definitions etc. expands. Each review and code list creation involves the agreement of an initial set of keyword search terms with clinical/domain experts, which are used within SAIL by the group to review against the content of the WRRS data, to find the items that best match the codes and code name descriptions as they exist in the WRRS. This is an iterative process, with the clinical/domain expertise being critical to process both at the initial and further stages to review and confirm the correct items are identified and then captured within the resulting code list definitions and data dictionary. Over time, where different interpretations or requirements may be based on study design as to what is agreed as appropriate for a given code list to be captured, different versions can be built and stored as per the requirements.

The specific mechanics of the process to define these research areas of interest are displayed in the flow diagram of the supplementary material (see Defining research area of interest in WRRS - Process flow)

Curating new definitions of interest

In order to curate new definitions of interest for research from the WRRS, it is necessary to clean and extract records (i.e. rows and columns) from the WRRS data provided to your SAIL project, which are relevant to your definition of interest.

Specific details on how to clean, extract and create working tables are detailed in the supplementary material under "Cleaning, extraction and insertion of WRRS data into working tables", which lists the relevant GitHub repositories.

Cleaning data

The VAL field in WRRS holds the test result (in the WRRS_OBSERVATION_RESULT table). When working with the data, it is advised and convenient to separate the VAL entries into numeric results only (VAL_NUMERIC) and text results only (VAL_TEXT) so that mathematical operations can be performed on the numeric results, and text results can be treated as such in isolation. This is something which we hope to be implemented in the research ready version of the data moving forward as standard.

Extracting data

An SQL code template is used to apply the research definition of interest. The template is modified by substituting in the codes from the particular definition in question. This is then run on the WRRS data.

Inserting data into working tables

Inserting data:

Due to the size of the WRRS data, it is not practical to select and insert results into the SAIL project data schema tables (known as the SAIL W schema), without first partitioning them into smaller sections (i.e. breaking them down into smaller chunks).

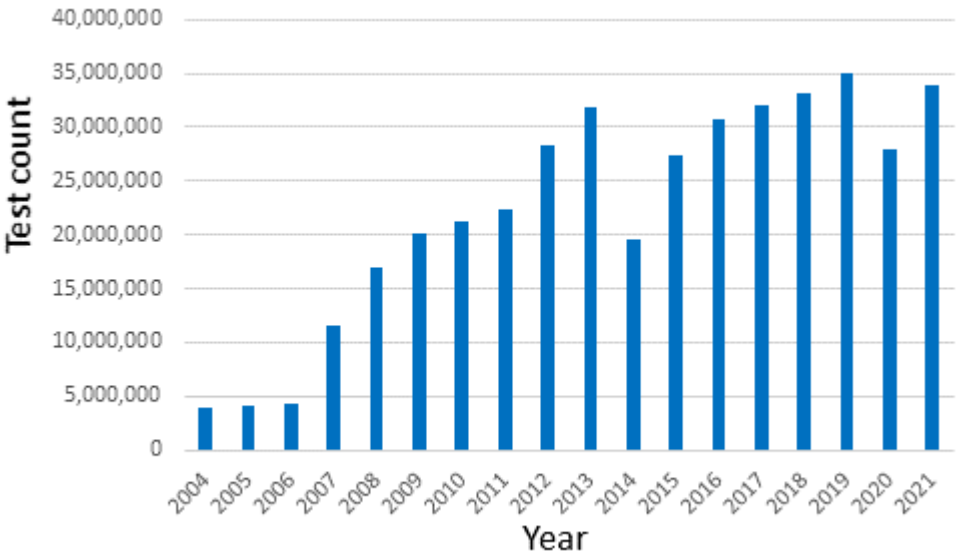
It is recommended to break results down into date sections, rather than person IDs (ALF), for the following reasons:

- It is not easy to identify the relative sizes of ALF partitions due to the non-intuitive nature of ALF values. It is easier and quicker to identify different years.
- When adding updated results, only the latest dates would need to be run if partitioning by date. Whereas the whole update would need to be redone if partitioning by ALF's.

Determining the size of partitions – an example:

Data was extracted for the Full Blood Count (FBC) test results between 2004 and 2021 inclusive (Figure 2). Firstly the date fields were examined per year to see the size of each data block:

Figure 2: Full Blood Count tests in WRRS 2004-2021



On performing the inserts, it was found that groupings of below 200 million were just about ok, but insert sizes of less than 100 million ran much more quickly – e.g. 2015-2018, 2019-2021.

In summary, in this example row counts of less than 100 million were preferable for the blocks of data to be inserted.

Units of measurement, reference ranges and result type

Units of measurement provided in the WRRS data are varied. Therefore, it is common practice to convert these to the most widely used unit of measurement for the particular category in question. Transformations may also be needed (e.g. converting 'per decilitre' to 'per litre'). Where this is relevant, the result (the VAL field) will likewise need to be transformed by multiplying by the required scaling factor.

Some results are associated with reference ranges. Where populated, these are likely to need harmonising similarly to the units of measurement. Harmonisation of ranges can also be applied to the values and the text entry display if this is deemed necessary. One approach to harmonising range values is to select the lowest low and the highest high to produce a broad encompassing range. However, the range selection method will be specific to the user's need, and clinician input is recommended.

Tables 1 and 2 show a selection of the unit of measurement and reference range entries for white blood count results before and after a harmonisation filter is applied.

Table 1: Units of measurement harmonisation

Units of measurement	
Before harmonisation	**After harmonisation**
*10^9/L	*10^12/L
/cmm	*10^6/L
^9/L	*10^9/L
^9/l	/CMM
cu/ mm	CU/MM
x10<[9]>/L	
x10E9/l	
x10^12/l	
x10^6/L	
x10^9/L	
x10^9/l	

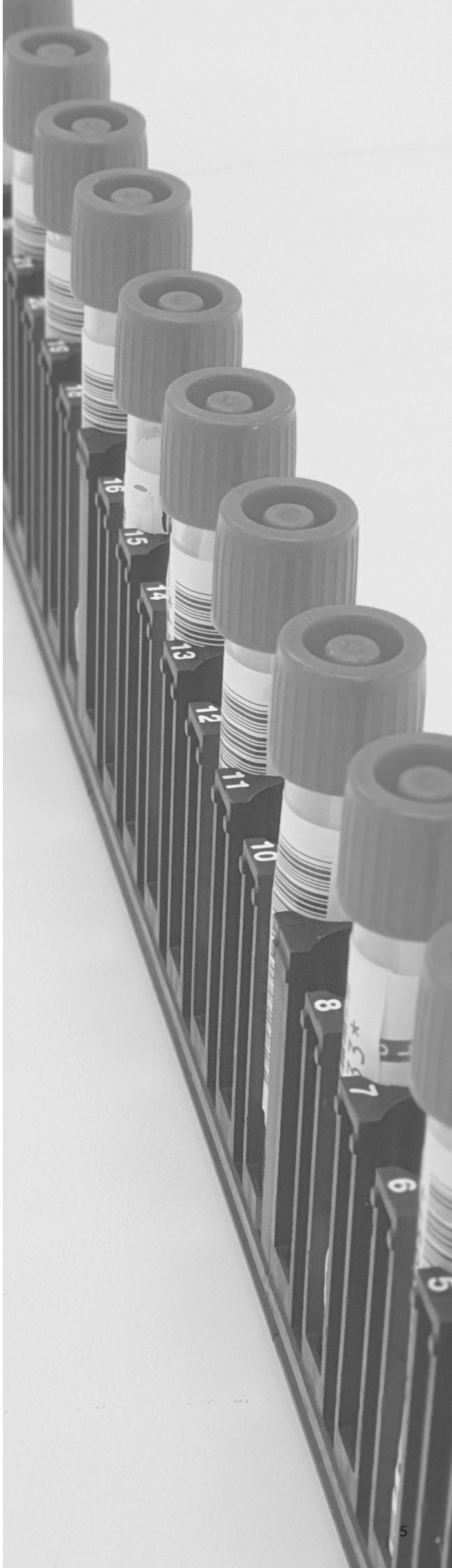


Table 2: Reference ranges harmonisation

Reference ranges		
Before harmonisation	**After harmonisation (of text)**	**After harmonisation (of values)**
4-11 4.0 - 11.0 4.0-11.0 5-15 5.0 - 15.0 5.0-15.0 Phone lab for ranges	4.0-11.0 5.0-15.0	4.0-15.0

The type of result entry is described in the VAL_TYPE field. These have been interpreted as described below and are presumptive descriptions based on HL7 data types [4] and guestimates from comparing to VAL entries.

Table 3: Result types with descriptions

Result type (VAL_TYPE)	
VAL_TYPE	**Description**
CE	coded element (seem to be text descriptions)
DT	date
ED	(all VAL are null) - presumably encapsulated data
MU	multiple value
NTE	Note (.nte file type)
RP	reference pointer - points to data on another system (e.g. URLs)
SN	structured numeric
ST	string
TM	time
TX	text (string data meant for user display on a screen)

These descriptions would need confirmation by clinical/domain expertise before they can be used in an overall implementation against the WRRS, so they should be taken as best guesses at this stage.

Redacted results

The results in WRRS are entered in the VAL field. As the original system that collects details in the NHS allows the entry of any information that the health care professional feels is relevant to the record, some of the result entries are in text form. As such, when the data are acquired into SAIL, a reduction process is completed to filter and remove any potentially disclosive results (e.g. Doctor's names etc.). Where this is the case, the text entry is replaced with {redacted}, rather than removed altogether, so that within SAIL, it can still be distinguished between redacted results and null entries. At the time of running this report (coverage until 31st December 2021), the percentages of VAL entry types in WRRS are shown in table 4:

Table 4: Percentages of redacted results in WRRS

Result entry type	Result count	% of total results
NULL	447,866,099	21.25
{redacted}	83,556,440	3.96
Available Data (i.e. not null not redacted)	1,576,287,998	74.79

WRRS Data Coverage

4,136,580 people were present in WRRS with coverage up to and including 31st December 2021, at the time of running this report. 3,977,814 of these people had data from 2004 onwards (using the authorised date in the WRRS_OBSERVATION_REQUEST table). 3,822,429 of these people were also found in the Welsh Demographic Service Data (WDSD) for records with a date valid from or following 2004 onwards.

For the purposes of our COVID-19 pandemic response work [5], an e-cohort has been created and maintained known as the C20-cohort, which consists of all people alive and known to the NHS in Wales, from the beginning of 2020. [6] Using this e-cohort as a population denominator of all potential people who are resident and alive in Wales and who have had an opportunity to interact with NHS Wales services since January 1st 2020, 76% of the people present in WRRS were able to be linked to the C20-cohort in SAIL.

Figures 3-5 examine the coverage of all (not just the C20-cohort) of the WRRS data in the form of person counts. A similar analysis can be performed for number of tests per date.

Dates are filtered to years 2004 onwards as data coverage prior to 2004 is minimal. The authorised date field (AUTHORISED_DTTM) in the WRRS_OBSERVATION_REQUEST table has been used to define dates here, although other date fields are available. Typically, the authorised date has fewer anomalous entries.

Person counts by (authorised) date are the number of people having one or more events within each distinct day. Person counts by month are the number of people having one or more events within each distinct day of the specified year/month. For example, if one person has multiple events on a particular day, that person is only counted once for that day.

Figure 3: WRRS person count 2004-2021 (30 day rolling average)

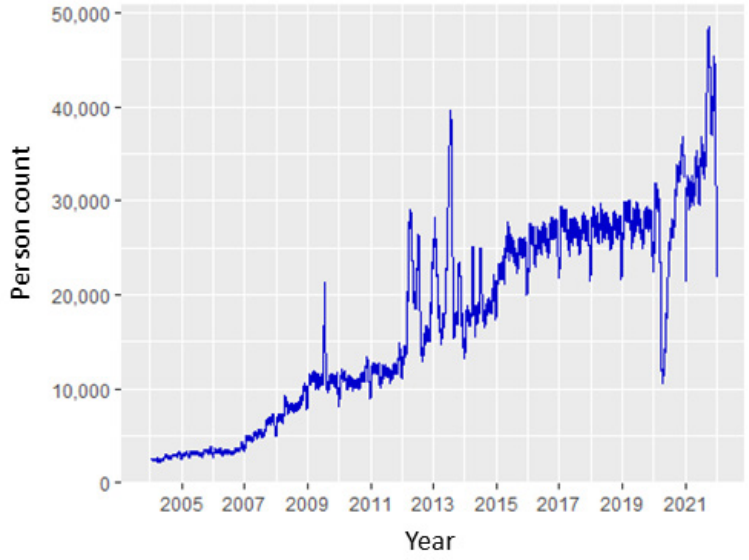


Figure 4: WRRS all test count 2004-2021 (30 day rolling average)

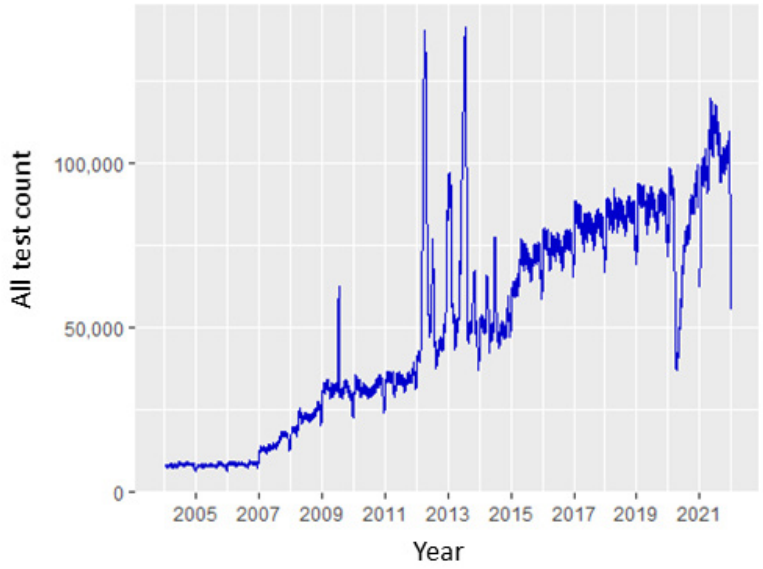
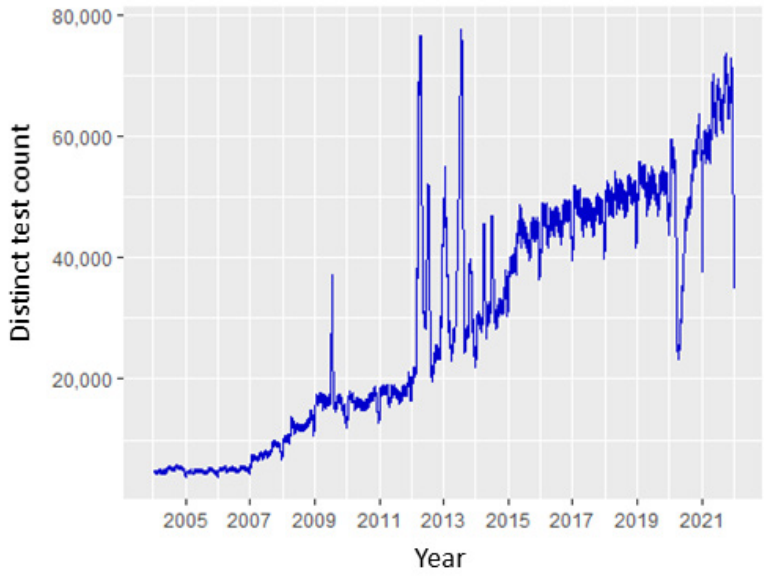


Figure 4: WRRS all test count 2004-2021 (30 day rolling average)



Automated coverage assessment algorithm

The data in WRRS is recorded and collected from different health boards across Wales, and the start periods vary as to when these systems and data began being used and collected. Each health board will have a start and end period between which we are confident there is a consistent flow of data. In order to assess these time periods, we look at counts of tests and people.

The quality of the data coverage is measured by taking the average count for a set time period (e.g. 7, 30, 90 days), then dividing the count at a particular time by the average. When the ratio drops below a set value (e.g. 0.3), the level of quality coverage period has ended. This is a way to mathematically describe the process of manually reviewing a plot of the data and deciding where the coverage is 'good'.

The Population Data Science group has developed an SQL algorithm to assess these levels of quality coverage periods. Figures 6 and 7 show the number of patients by health board and the number of tests by health board. A log10 scale has been used due to the wide variation between health boards.

Figure 6: Health Board coverage by person count (LOG 10 scale) - 2004-2021

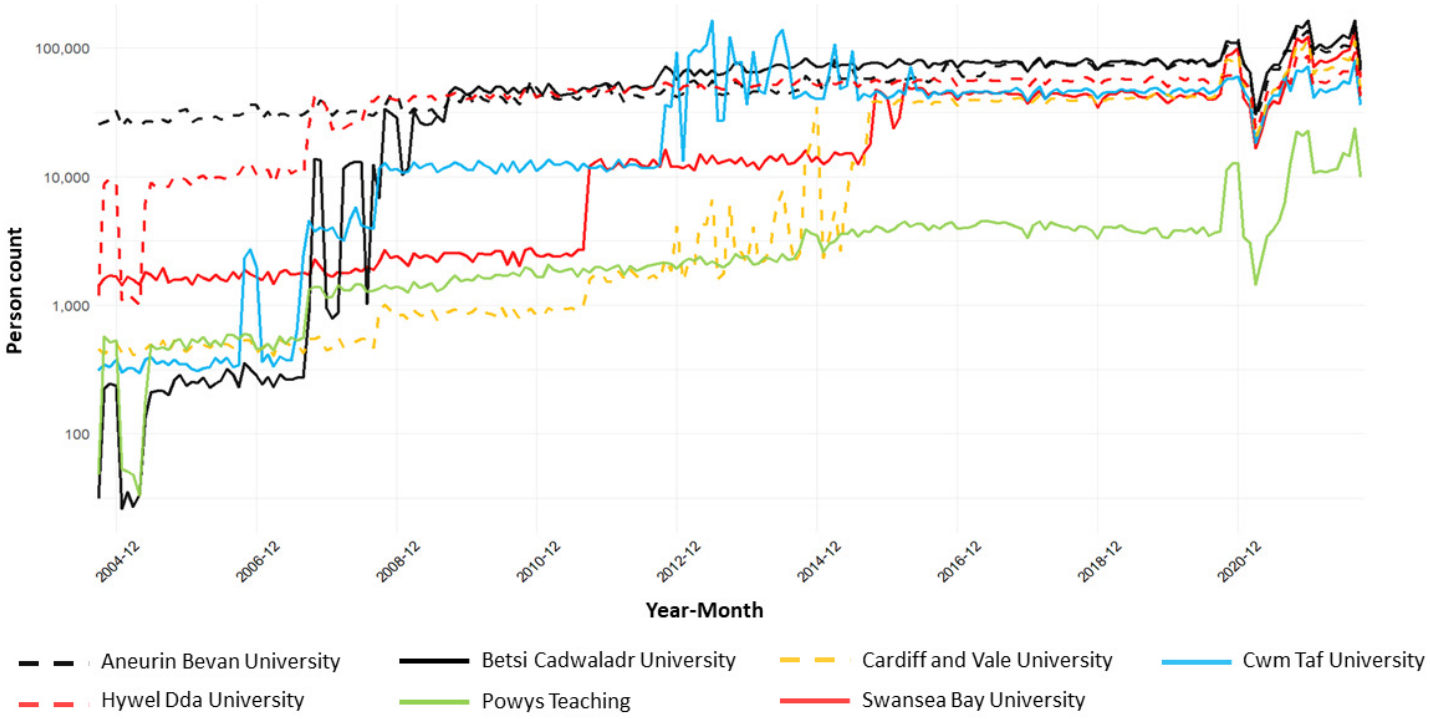
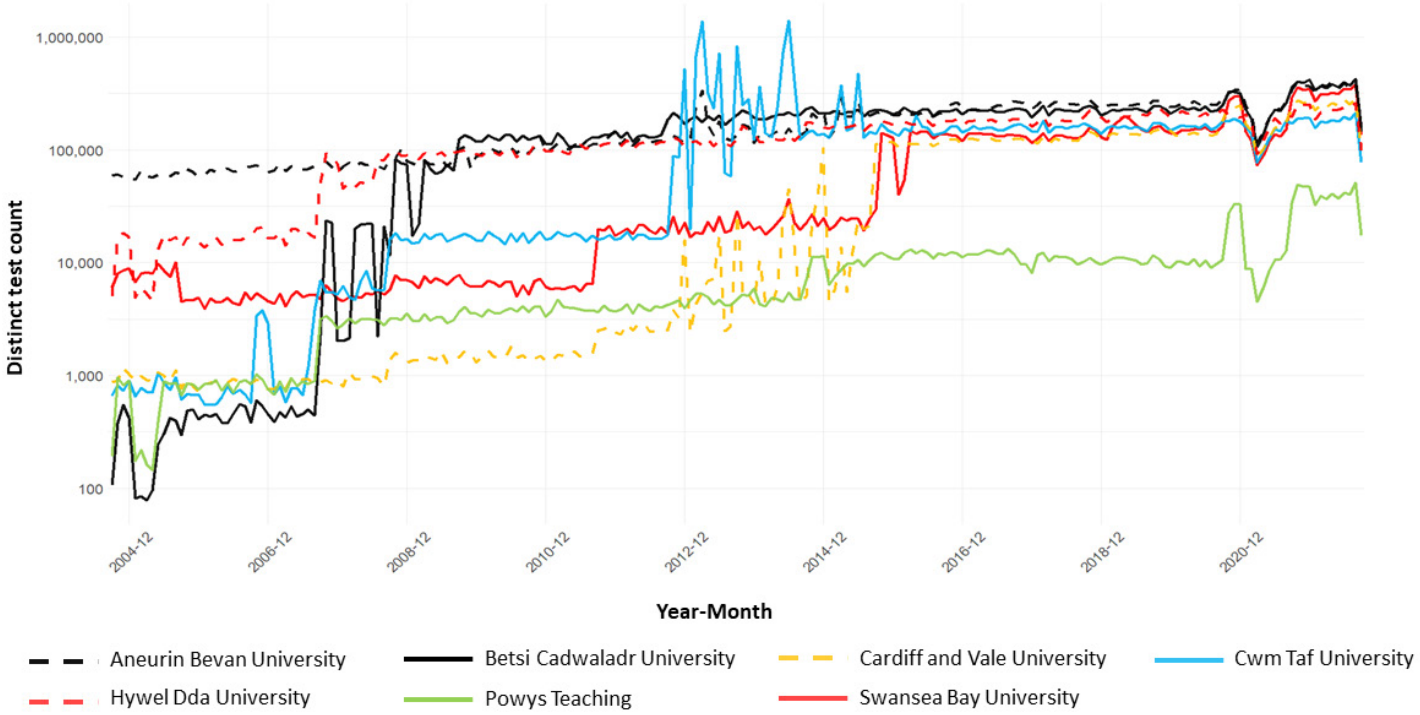


Figure 7: Health Board coverage by distinct test count (LOG 10 scale) - 2004-2021



Optimum parameters were determined as ratio to average = 0.3, time period = 30 days. Using this automated approach to assessing levels of quality of coverage data for each health board, start and end dates of coverage were assessed as:

Table 5: Person count coverage by health board as assessed by automated algorithm

Health Board	Count type	Start date	End date
All	Person count	2008-02-14	2021-12-31
Aneurin Bevan University	Person count	2004-02-05	2021-12-31
Betsi Cadwaladr University	Person count	2008-04-14	2021-12-31
Cardiff and Vale University	Person count	2014-07-12	2021-12-31
Cwm Taf University	Person count	2012-03-24	2021-12-31
Hywel Dda University	Person count	2007-01-20	2021-12-31
Powys Teaching	Person count	2007-01-20	2021-12-31
Swansea Bay University	Person count	2012-05-23	2021-12-31

Table 6: Test date count coverage by health board as assessed by automated algorithm

Health Board	Count type	Start date	End date
All	Test date count	2008-02-14	2021-12-31
Aneurin Bevan University	Test date count	2004-02-05	2021-12-31
Betsi Cadwaladr University	Test date count	2008-05-14	2021-12-31
Cardiff and Vale University	Test date count	2014-12-09	2021-12-31
Cwm Taf University	Test date count	2012-03-24	2021-12-31
Hywel Dda University	Test date count	2007-01-20	2021-12-31
Powys Teaching	Test date count	2007-02-19	2021-12-31
Swansea Bay University	Test date count	2015-02-07	2021-12-31

The STATS_CURR_CENSUS_LSOA_CD field in WRRS_OBSERVATION_REQUEST was used to differentiate health boards.

Conclusion

The WRRS data source enables the potential to provide great impact in various potential research questions and projects that were previously not possible without access to these types of data. It is one of the first of its kind, in that it is available and updated regularly and is linkable other population-scale data within the SAIL Databank. Further work on understanding and harmonising the contents opens up options for anonymised individual-level health care utilisation and outcomes not possible from other electronic health record data sources currently. With the considerations we have outlined around the very large and complex nature of the data, we see the utility and strengths of these data being used as part of ongoing research and intelligence in Wales.

Glossary

Term	Description
Algorithm	A set of rules to be followed
Biomarker	Measurable medical characteristics
Fields	The individual columns of a data table
Full Blood Count (FBC)	A test to check the types and numbers of cells in blood
GIT	Version control software to track changes in a set of files
HCP	Health Care Professional
LSOA	Lower-layer super output area: 1-3K people, or 400-1,200 households
Metadata	Data that provides information about other data, but not the content
Parameter	A number forming part of the definition of an allowable range
Pathology	In this context: The analysis of tissue, cell and body fluid samples
Platelet Count	A test to measure the number of platelets in blood (subcategory of FBC)
SAIL	Secure Anonymised Information Linkage (see www.saildatabank.com)
Schema	Organisational structure of a database, defining the table storage
SQL	Structured Query Language – programming language for data manipulation

References

[1] https://www.wales.nhs.uk/news/43900 (accessed 10/11/2021)

[2] https://dhcw.nhs.wales/systems-and-services/secondary-care/welsh-laboratory-information-management-system/ (accessed 17/11/2021)

[3] https://web.www.healthdatagateway.org/dataset/71d37610-ac55-432d-82a3-bdb04407acd8 (accessed 10/2/2022)

[4] https://www.hl7.org/ (accessed 10/2/2022)

[5] https://popdatasci.swan.ac.uk/news/one-wales/ (accessed 10/2/2022)

[6] Lyons J, Akbari A, Torabi F, et al. Understanding and responding to COVID-19 in Wales: protocol for a privacy-protecting data platform for enhanced epidemiology and evaluation of interventions. BMJ Open 2020;10:e043010. doi:10.1136/bmjopen-2020-043010

Supplementary Material

WRRS metadata

Table name: WRRS_OBSERVATION_REQUEST

Field Name	Type	Description
ALF_E	INTEGER	LINKAGE FIELD - Person identifier (encrypted NHS number)
ALF_STS_CD	INTEGER	ALF status code
ALF_MTCH_PCT	DOUBLE	ALF match percentage (informs ALF_STS_CD)
SPCM_COLLECTED_DT	DATE	Specimen collected date (need to confirm whether identical to the same field in report table)
ID_E	BIGINT	Request ID (encrypted)
REPORT_SEQ	INTEGER	LINKAGE FIELD - Report sequence
REQUEST_SEQ	INTEGER	LINKAGE FIELD - Request sequence
CD	VARCHAR	Code of the request (some similarities to CODE in result table, but not identical)
NAME	VARCHAR	Name of the request (similar but not identical to NAME in result table)
PROV_SYSTEM_CD	VARCHAR	Provider system code (very similar to NAME in this table, but not always identical)
READ_CD	VARCHAR	Read code of request
OBSERVATION_STS_CD	VARCHAR	Observation status code
SENSITIVITY_STS_CD	VARCHAR	Sensitivity status code
AUTHORISEDBYID	INTEGER	ID of authoriser
REPORTDISPLAYORDER	VARCHAR	Report display order
AUTHORISED_DTTM	TIMESTAMP	Date and time was authorised
AVAIL_FROM_DT	DATE	Date of data upload into SAIL

Table name: WRRS_OBSERVATION_RESULT

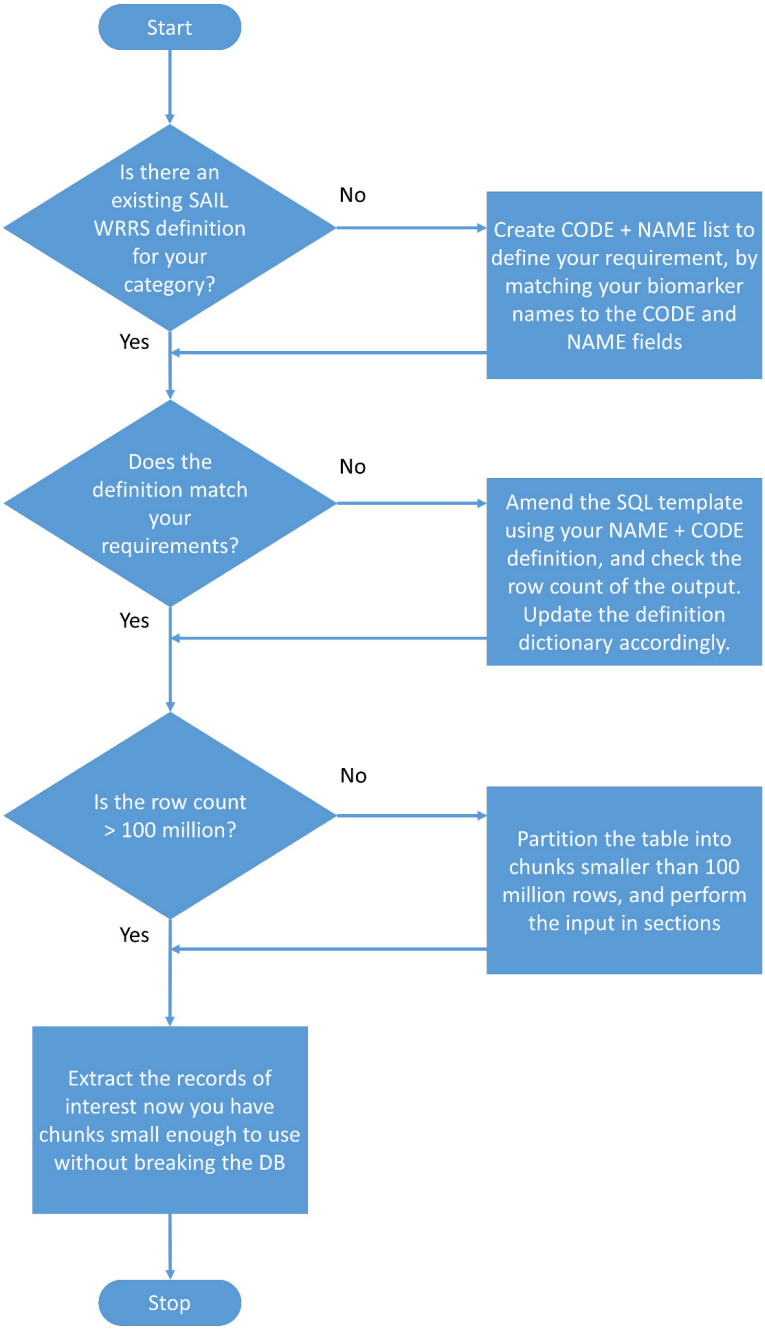
Field Name	Type	Description
ALF_E	INTEGER	LINKAGE FIELD - Person identifier (encrypted NHS number)
ALF_STS_CD	INTEGER	ALF status code
ALF_MTCH_PCT	DOUBLE	ALF match percentage
SPCM_COLLECTED_DT	DATE	Specimen collected date
OBSERVATION_REQST_ID_E	BIGINT	Observation request ID (encrypted)
REQUEST_SEQ	INTEGER	LINKAGE FIELD - Request sequence
REPORT_SEQ	INTEGER	LINKAGE FIELD - Report sequence
VAL_TYPE	VARCHAR	Test result value type (VAL entry type)
CODE	VARCHAR	Code of the test
NAME	VARCHAR	Name of the test
PROV_SYSTEM_CD	VARCHAR	Provider system code
READ_CD	VARCHAR	Read code of test
VAL	VARCHAR	Test result value
UNITOFMEASUREMENT	VARCHAR	Test result unit
REFERENCERANGE	VARCHAR	Test result reference range
ABNORMAL_STS_CD	VARCHAR	Whether the result (VAL) lies within normal range
AVAIL_FROM_DT	DATE	Date of data upload into SAIL

Table name: WRRS_REPORT

Field Name	Type	Description
ID_E	BIGINT	Report ID (encrypted)
REPORT_SEQ	INTEGER	LINKAGE FIELD - Report sequence
MASTERREPORTID_E	BIGINT	Master report ID (encrypted)
REPORTVERSION	INTEGER	Report version
REPORTTITLE	VARCHAR	Report title
PROV_TYPE_CD	VARCHAR	Provider type code
REPORTSENSITIVITY_STS_CD	VARCHAR	Report sensitivity status code
REPORTSENSITIVE	VARCHAR	Flag as to whether the report is of sensitive nature (true/false)
ABNORMALRESULTS	VARCHAR	Abnormal result flag (Y/N)
PLACER_REQST_NUM	INTEGER	Placer request number
PATHWAY_ID	INTEGER	Pathway ID
PROV_SYSTEMID_E	BIGINT	Provider system ID (encrypted)
PROV_DEPT_ID_E	BIGINT	Provider department ID (encrypted)
PROV_DEPT_SITEID	INTEGER	Provider department site ID (lookup table = SAILWRRSREFV.WRRS_REFERENCE)
PROV_DEPT_ORG_ID_E	BIGINT	Provider department organisation ID (encrypted)
SUBJECT_LOC_ID	INTEGER	Subject location ID
SUBJECTSITEID	INTEGER	Subject site ID (lookup table = SAILWRRSREFV.WRRS_REFERENCE)
SUBJECT_ORG_ID_E	BIGINT	Subject organisation ID (encrypted)
REQUESTORID_E	BIGINT	Requestor ID (encrypted)
REQUESTOR_SPEC_ID_E	BIGINT	Requestor speciality ID (encrypted)
PLACERID_E	BIGINT	Placer ID (encrypted)
PAT_TYPE_ID	INTEGER	Patient type ID
PAT_CAT_ID	INTEGER	Patient category ID
REPORT_DTTM	TIMESTAMP	Report date and time
SPCM_COLLECTED_DT	DATE	Specimen collected date
SPCM_COLLECTED_TM	VARCHAR	Specimen collected time
SPCM_RECEIVED_DT	DATE	Specimen received date
SPCM_RECEIVED_TM	VARCHAR	Specimen received time
UPDATED_DTTM	TIMESTAMP	Updated date and time
ISARCHIVED	INTEGER	Archived flag
TRANSFORMATIONPROCESSID	INTEGER	Transformation process ID
ISBINARY	VARCHAR	Is binary (true/false)
ALF_E	INTEGER	LINKAGE FIELD - Person identifier (encrypted NHS number)
ALF_STS_CD	INTEGER	ALF status code
ALF_MTCH_PCT	DOUBLE	ALF match percentage (informs ALF_STS_CD)
STATS_CURR_CENSUS_LSOA_CD	VARCHAR	LSOA code from current census
AVAIL_FROM_DT	DATE	Date of data upload into SAIL

Defining research area of interest in WRRS - Process flow

Figure 8: Defining research definition of interest in WRRS - Process flow



Defining research area of interest in WRRS – Definitions

The definitions derived in WRRS are listed in the document WRRS_RRDA_definitions_dictionary.docx, which is held on GitHub repository (https://github.com/SwanseaUniversityMedical/WRRS) and also within SAIL. This consists of the code and name field lists which have been decided upon in order to define research areas of interest. As users derive new definitions or improve existing definitions, they should update this document with the new/improved definitions.

Cleaning, extraction and insertion of WRRS data into working tables for research

A working SQL template to extract definitions is available in SAIL on request, as well as at the GitHub repository listed above. Details on the template and how to apply it can be found on the repository and supporting documentation.

The template is based on the FBC definition, and will need to be modified to suit the required NAME + CODE definition.

If you are interested in accessing the list of currently available defintions and/or collaborating on further developments in the future please get in contact with the team via email on @swansea.ac.uk

This Data Explained has been created by a research team working closely with the WRRS data. This document is intended as a starting point for researchers who are looking to use the WRRS data as part of an approved data linkage research project and is correct at the point of publication.

ADR Wales does not hold datasets. This data is held securely and anonymously within the SAIL Databank. Subject to safeguards and approvals, data held within the SAIL Databank can be linked together to address important research questions. This advanced data linkage research platform is the UK's first single resource for population, health and social care data intended solely for research. A list of the anonymous data held within the SAIL Databank can be found here: https://data.ukserp.ac.uk/Organisation/Index/0

Produced by ADR Wales

Website: ADRUK.org
Twitter: @ADR_Wales

ADR Wales brings together specialist teams, data science experts, and statisticians as part of the Economic and Social Research Council (part of UK Research and Innovation) funded ADR UK. Our team is made up of specialists in their field from Swansea University Medical School, the Wales Institute of Social and Economic Research, Data and Methods (WISERD) at Cardiff University and the SAIL Databank at Swansea University with statisticians, economists and social researchers from Welsh Government. Together ADR Wales develops new evidence which supports the Welsh Government's national strategy, Prosperity for All to improve the lives of people in Wales.

For further information please contact
Cathrine.E.Richards@Swansea.ac.uk

ADR UK
Data-driven change

UKRI Economic and Social Research Council

ADR Wales Partners

Cardiff University
Prifysgol Caerdydd

Swansea University
Prifysgol Abertawe
Medical School
Ysgol Feddygaeth

Llywodraeth Cymru
Welsh Government

SAIL DATABANK

WISERD
Sefydliad Ymchwil Cymdeithasol ac Economaidd a Data Cymru
Wales Institute of Social and Economic Research and Data