**- Data Brief -**

**The Welsh Results Reports Service (WRRS) data**

**Wales population-scale pathology data, a national data asset.**

**Authors: Gareth Davies, Ashley Akbari.**

**Date version generated: 24th May 2022**

**Index**

**Introduction 2**

-Data location 2

-Improvements to services 2

**How the pathology data are collected 2**

**What the pathology data can be used for 3**

**Defining research areas of interest in WRRS 3**

**Curating new definitions of interest 4**

-Cleaning data 4

-Extracting data 4

-Inserting data into working tables 4

-Units of measurement, reference ranges and result type 6

**WRRS Data Coverage 7**

-Automated coverage assessment algorithm 9

**Conclusion - so what? 10**

**Glossary 11**

**References 12**

**Supplementary Material 13**

-WRRS metadata 13

-Defining research area of interest in WRRS - Process flow 16

-Defining research area of interest in WRRS – Definitions 16

-Cleaning, extraction and insertion of WRRS data into working tables for research 17

**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]:

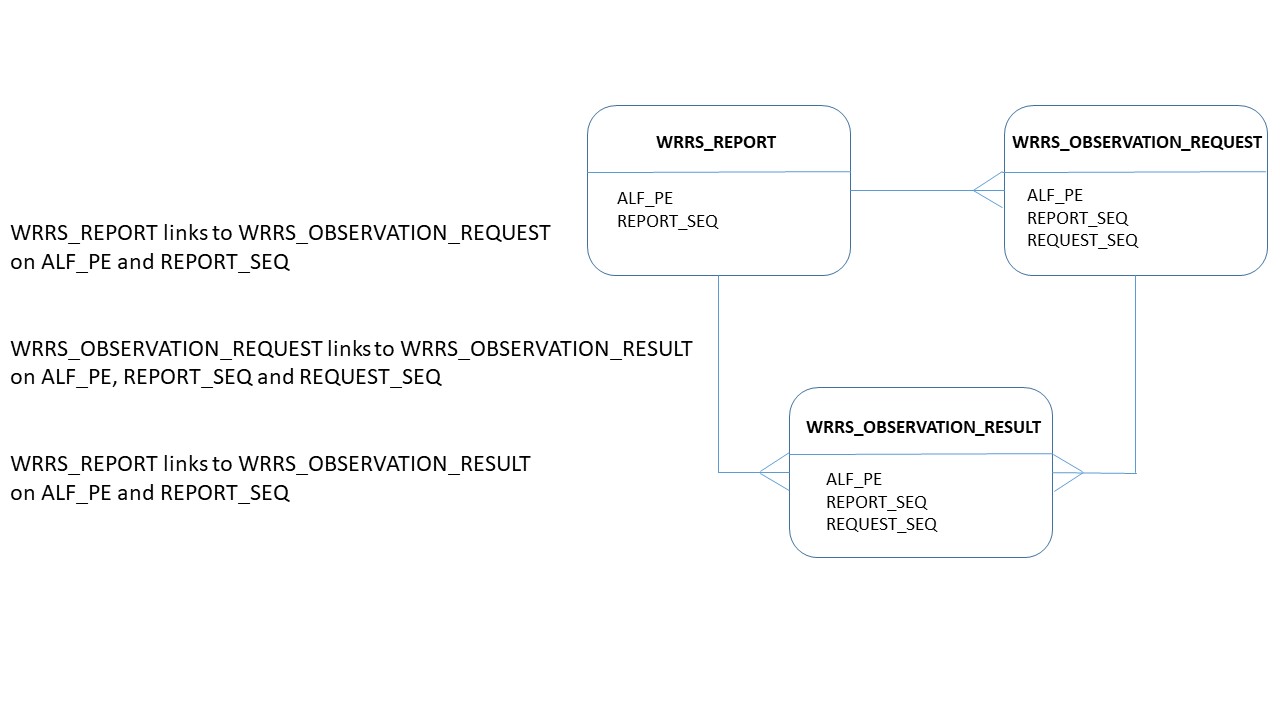
* Improved clinical governance.
* Flexibility, portability and adaptability of service provision.
* Improved demand management and forward planning.
* An improved patient experience.
* Service standardisation.
* Provision of information to the Welsh general practice record.

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/ HDR innovation gateway [3].

**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 a 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.

***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).

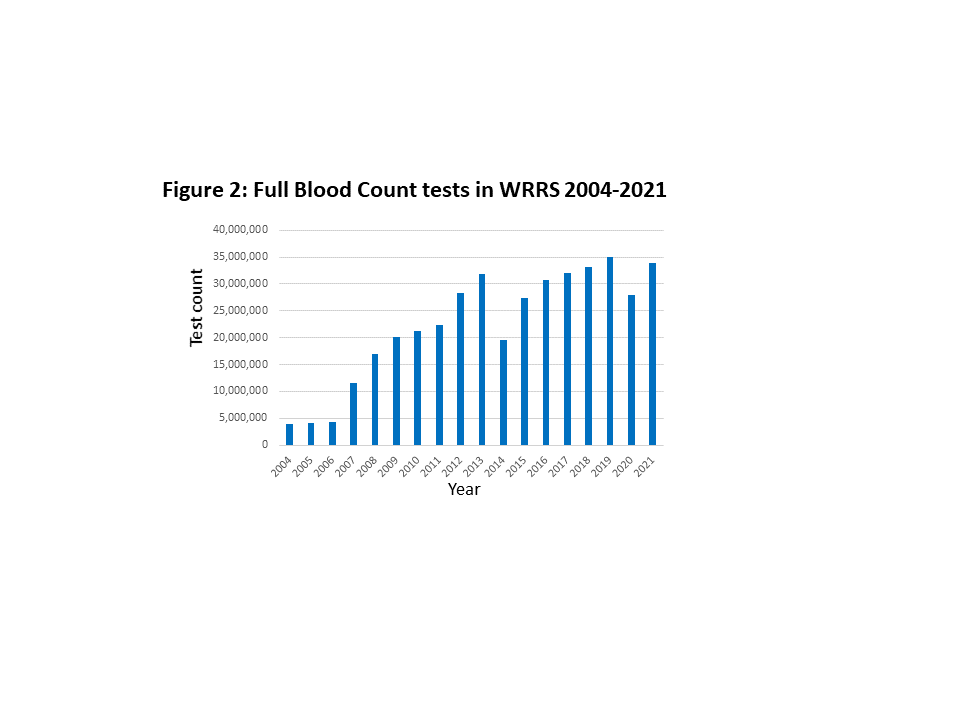
It is recommended to break results down into date sections, rather than person IDs (ALF), for the following reasons:

• 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:



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, row counts of less than 100 million are 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 | 4.0-15.0 |
| 4.0 - 11.0 | 5.0-15.0 |  |
| 4.0-11.0 |  |  |
| 5-15 |  |  |
| 5.0 - 15.0 |  |  |
| 5.0-15.0 |  |  |
| Phone lab for ranges |  |  |

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 30th April 2022), 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 | 290,258,402 | 15.74 |
| {redacted} | 77,584,491 | 4.21 |
| not null not redacted | 1,476,052,375 | 80.05 |

**WRRS Data Coverage**

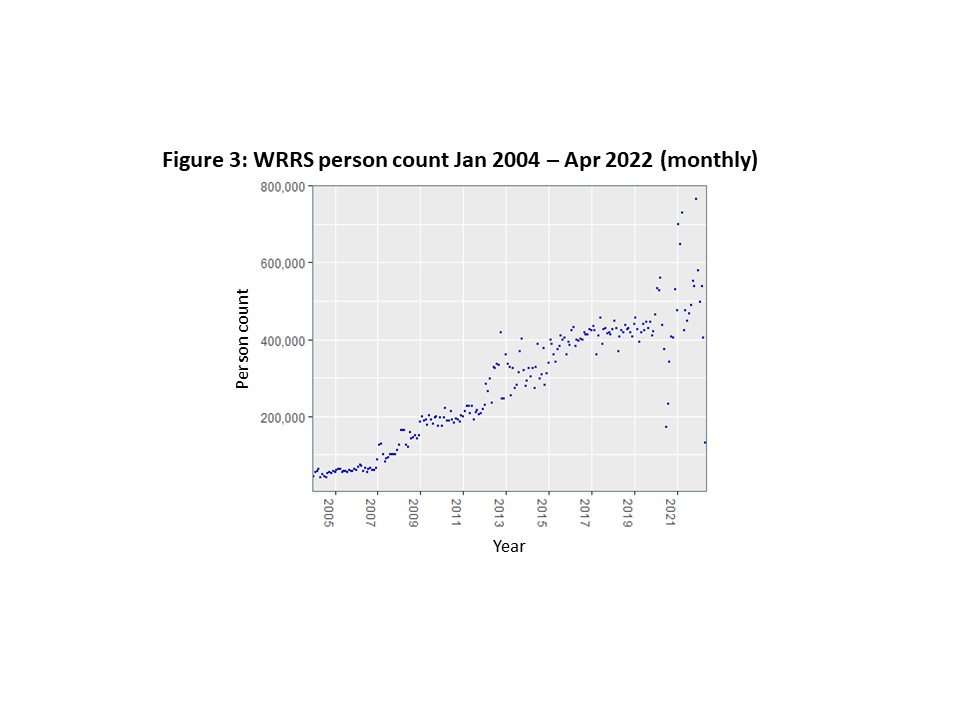
4,126,810 people were present in WRRS with coverage up to and including 30th April 2022, at the time of running this report. 4,042,865 of these people had data from 2004 onwards (using authorised date in the WRRS\_OBSERVATION\_REQUEST table). 3,883,516 of these people were also found in the Welsh Demographic Service Data (WDSD) for dates 2004 onwards (WDSD FROM\_DT >=2004 or WDSD TO\_DT>=2004).

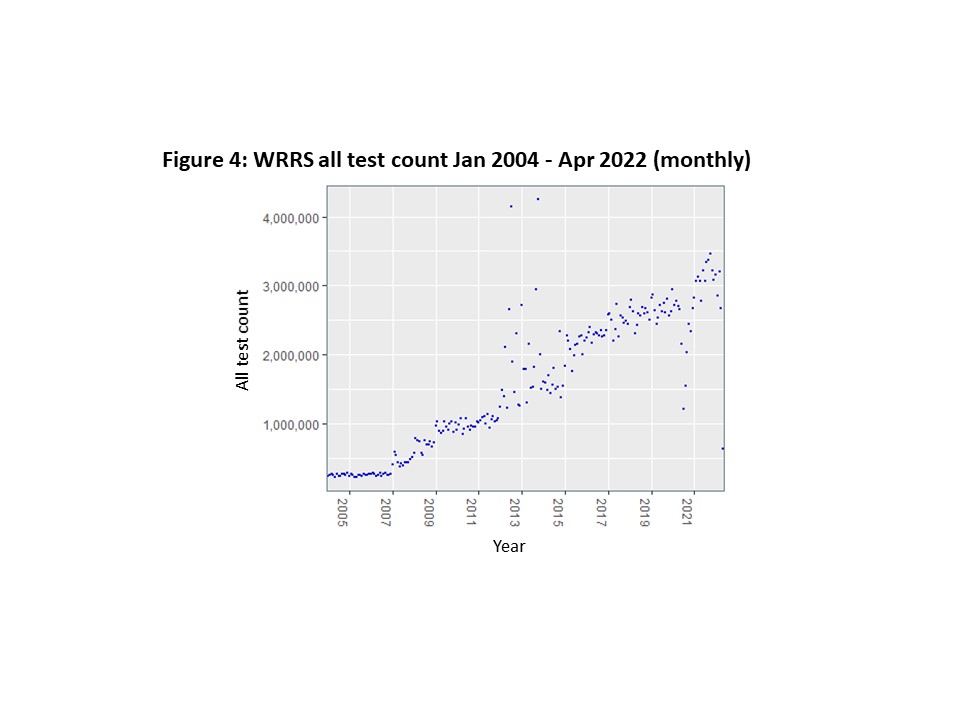
For the purposes of our COVID-19 pandemic response work [5], an e-cohort has been created and maintained known as the C20-cohort, which consists of all people alive and known to the NHS in Wales, from the beginning of 2020. [6] Using this e-cohort as a population denominator of all potential people who are resident and alive in Wales and who have had an opportunity to interact with NHS Wales services since January 1st 2020, 77% of the people present in WRRS were able to be linked to the C20-cohort in SAIL.

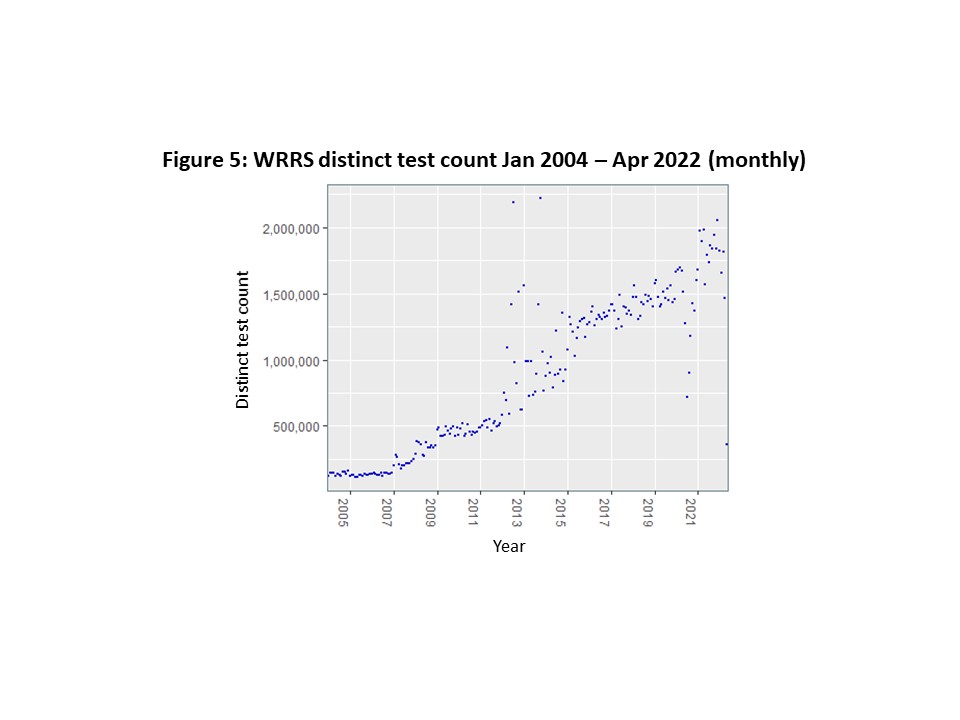
Figures 3-5 examine the coverage of all (not just C20-cohort) WRRS data, in the form of person counts and counts of number of tests per date.

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 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.

****



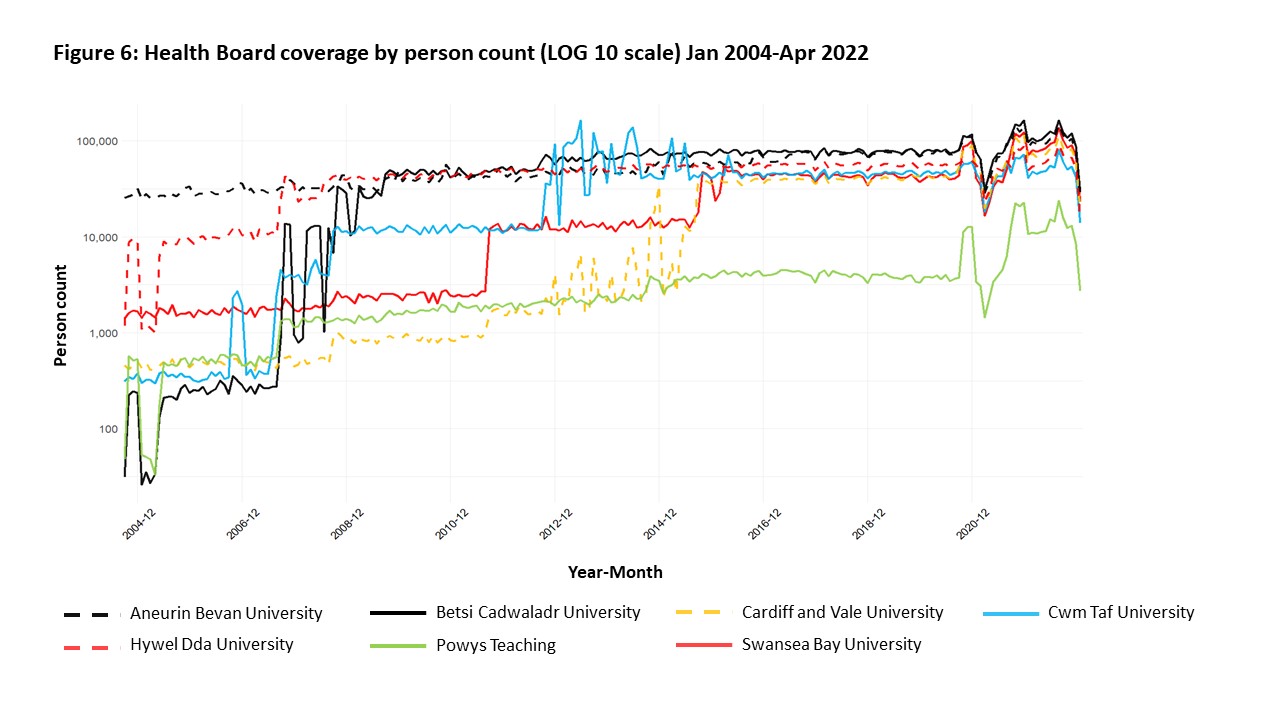


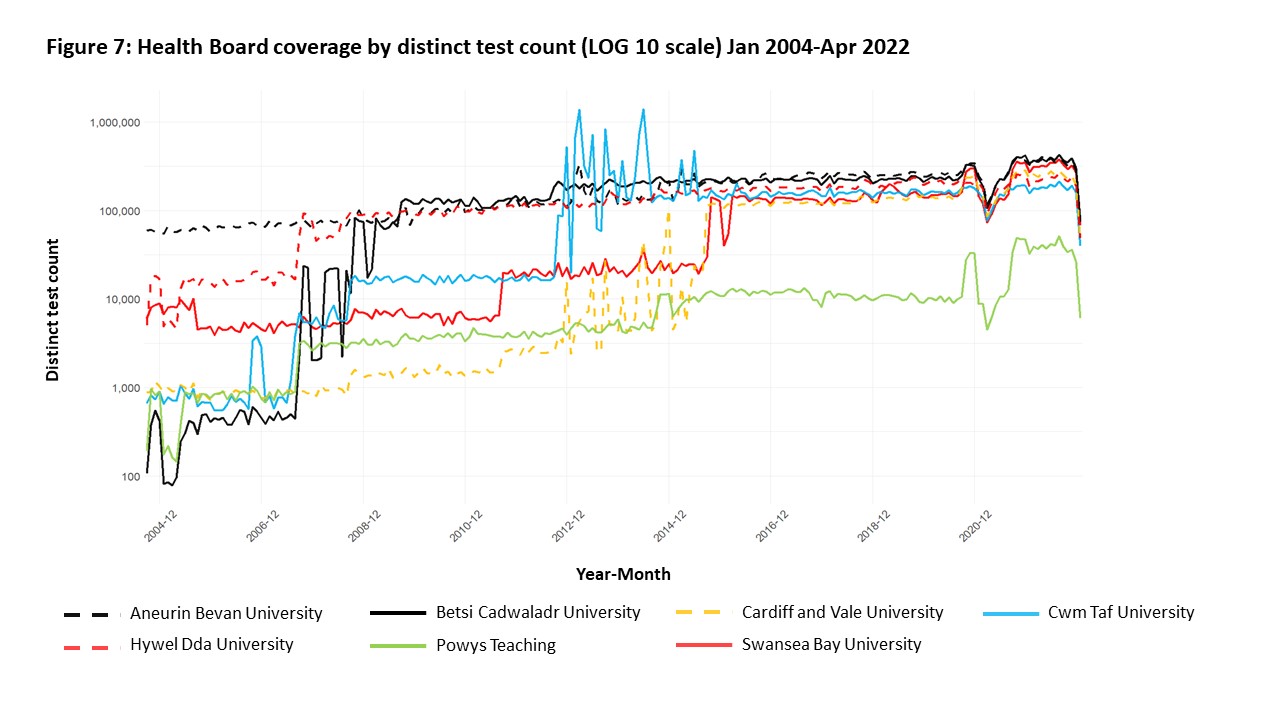
***Automated coverage assessment algorithm***

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'.

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





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 | 2022-04-30 |
| Aneurin Bevan University | Person count | 2004-02-05 | 2022-04-30 |
| Betsi Cadwaladr University | Person count | 2008-04-14 | 2022-04-30 |
| Cardiff and Vale University | Person count | 2014-07-12 | 2022-04-30 |
| Cwm Taf University | Person count | 2012-03-24 | 2022-04-30 |
| Hywel Dda University | Person count | 2007-01-20 | 2022-04-30 |
| Powys Teaching | Person count | 2007-01-20 | 2022-04-30 |
| Swansea Bay University | Person count | 2012-05-23 | 2022-04-30 |

**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 | 2022-04-30 |
| Aneurin Bevan University | Test date count | 2004-02-05 | 2022-04-30 |
| Betsi Cadwaladr University | Test date count | 2008-05-14 | 2022-04-30 |
| Cardiff and Vale University | Test date count | 2014-12-09 | 2022-04-30 |
| Cwm Taf University | Test date count | 2012-03-24 | 2022-04-30 |
| Hywel Dda University | Test date count | 2007-01-20 | 2022-04-30 |
| Powys Teaching | Test date count | 2007-02-19 | 2022-04-30 |
| Swansea Bay University | Test date count | 2015-02-07 | 2022-04-30 |

The STATS\_CURR\_CENSUS\_LSOA\_CD field in WRRS\_OBSERVATION\_REQUEST was used to differentiate health boards.

**Conclusion - so what?**

The WRRS dataset is very large and complex, but holds the potential to provide great impact in research. It is one of the first of it’s kind in that it is available to be linked to a nation wide population. Along with the consideratins we have laid out here, it is 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 |

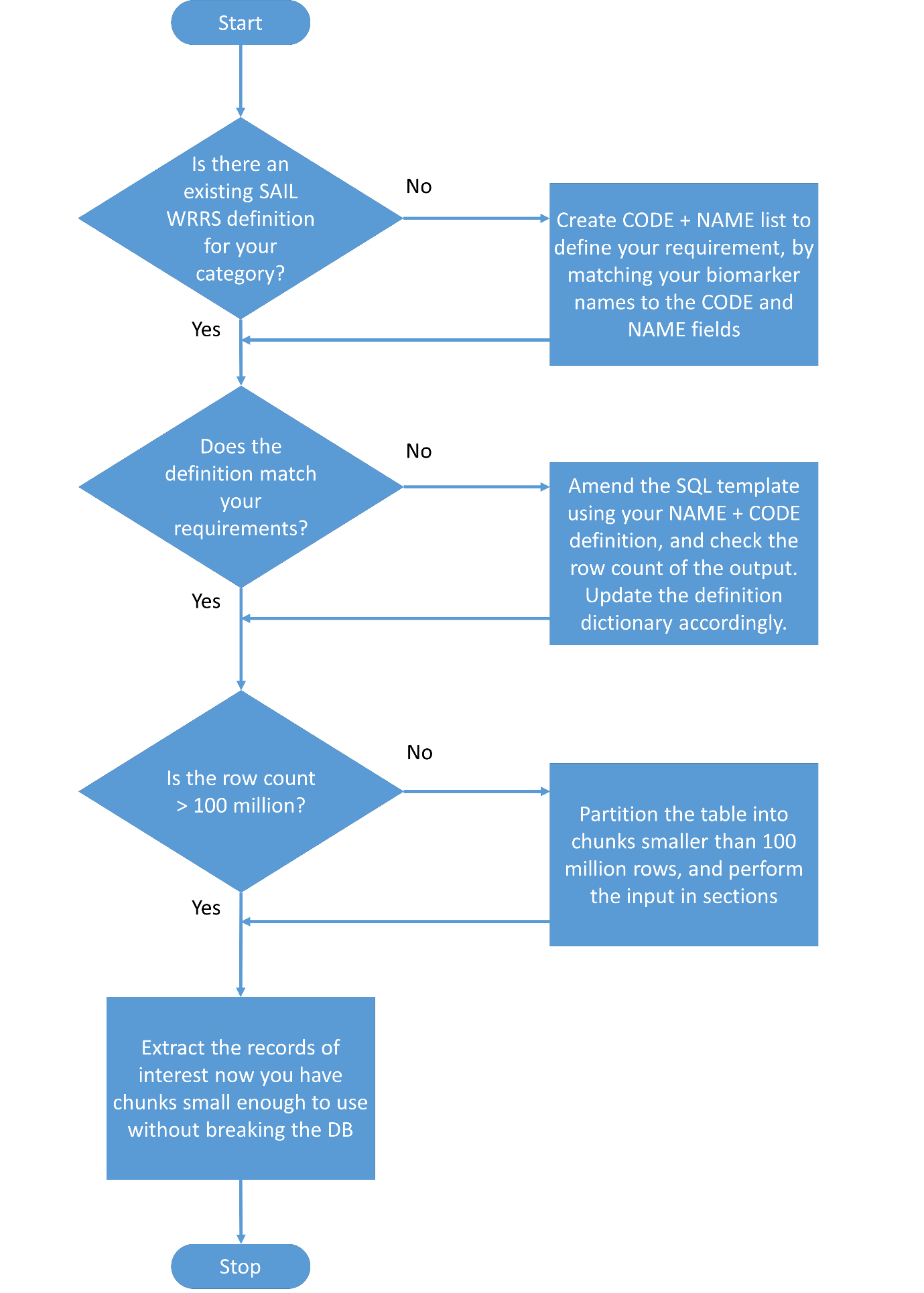
Table name: WRRS\_OBSERVATION\_RESULT

|  |  |  |
| --- | --- | --- |
| **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 |

***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*