International Journal of Population Data Science

Journal Website: www.ijpds.org

# Adapting historical clinical genetic test records for anonymised data linkage: obstacles and opportunities

Robert T. Maddison[1], Karen R. Reed[2], Rebecca Cannings-John[3], Fiona Lugg-Widger[3], Thomas Stoneman[4], Sarah Anderson[4], and Andrew E. Fry[1,4]

Submission History

| | |
|---|---|
| Submitted: | 13/11/2024 |
| Accepted: | 10/12/2024 |
| Published: | 28/02/2025 |

[1]Wales Gene Park, Division of Cancer and Genetics, Cardiff University, Canolfan Iechyd Genomig Cymru/Wales Genomic Health Centre, Cardiff Edge Business Park, Longwood Drive, Whitchurch, Cardiff, CF14 7YU, UK
[2]Centre for Medical Education, Cardiff University, Cardiff, CF14 4XN, UK
[3]Centre for Trials Research, Cardiff University, Cardiff, CF14 4XN, UK
[4]All-Wales Medical Genomics Service, Canolfan Iechyd Genomig Cymru / Wales Genomic Health Centre, Cardiff Edge Business Park, Longwood Drive, Whitchurch, Cardiff, CF14 7YU, UK

## Abstract

### Introduction

Cystic fibrosis (CF) heterozygotes (also known as 'carriers') are people who have one mutated copy of the *CFTR* gene. Research into the health risks of CF carriers has been limited by a lack of large cohorts tested for CF carrier status, but routine clinical testing identifies CF carriers in the population. Such test records additionally contain large amounts of clinical information, making them a valuable research resource to not only identify CF carriers in the population but also to provide additional data not found elsewhere.

### Methods

Following governance approvals, we adapted 30 years worth of CF genetic testing records generated by the All-Wales Medical Genomics Service (AWMGS) and submitted them to the SAIL Databank for anonymised linkage.

### Results

Unexpected obstacles meant that a minimum amount of clinical information could be annotated ahead of linkage. The raw data were highly heterogeneous due to the records' longitudinal collection and clinical origins, making standardisation difficult. Moreover, the presence of unique identifiers in the clinical data violated the separation principle, requiring manual annotation to produce a cleaned dataset. Explicit identification of patients or their relatives throughout the records complicated split file anonymisation.

### Conclusion

Extracting useful information from historical clinical genetic test records is a significant challenge with technical and governance aspects. The mixing of unique identifiers with clinical data in heterogeneous, unstructured free text combined with a lack of automated tools meant that manual annotation was required to adhere to the separation principle. As such, only a minimum of the available clinical data was annotatable within the project timeline and mutually exclusive access to the identifiable and pseudonymised data meant that annotations could not later be validated. Future efforts to link clinical genetic test records for research must consider these challenges in their approach.

### Keywords

data linkage; historical data; cystic fibrosis

*Corresponding Author:
*Email Address:* maddisonr@cardiff.ac.uk (Robert T. Maddison)

https://doi.org/10.23889/ijpds.v10i1.2924

# Introduction

Comprehensive data linkage infrastructure supports delivery of research that is of significant benefit to public health. Understanding the challenges associated with unlocking new data flows is key to securing the resources required for driving future research. Cystic fibrosis (CF) is a life-limiting genetic disorder caused by inheriting two pathogenic variants in the cystic fibrosis transmembrane conductance regulator (*CFTR*) gene. The CF Registry is a curated data product that captures longitudinal clinical information for CF patients in the UK, and has been utilised in data linkage environments to enhance the accuracy of CF research [1]. However, routinely collected clinical data, such as from genetics services, is not commonly utilised for research.

The All-Wales Medical Genomics Service (AWMGS) has provided *CFTR* mutation analysis for diagnostic purposes and newborn screening for 30 years. These reports represent a rich resource of information about people with CF (pwCF), CF carriers, and individuals who received a negative result (i.e. no mutations found), including genotypes, family history, and clinical indications for testing. Automated approaches to converting 'messy' historical records into structured analytical-ready formats are an active area of development [2] but this requires case-by-case design to accommodate the specific dataset being cleaned.

Here, we describe efforts to adapt 30 years' of clinical reports pertaining to CF tests generated by the AWMGS (the AWMGS CF Test Record) for anonymised data linkage project within the Secure Anonymised Information Linkage Databank (SAIL) Databank [3, 4].

# Description of data source

## Data contents

A total of 16,181 individual CF gene test records, collected between 1987 and 2023, are stored at AWMGS in an SQL compatible laboratory information management system (LIMS). Early records (pre-1995) are typically incomplete. The longitudinal nature of these records means their contents vary widely in formatting and scope, but broadly encompass three major areas of information: who the patient is, why they have been referred, and the result of the test.

Personal identifiable information (PII) includes forename, surname, NHS number, date of birth, and postcode at the time of testing, but not gender. Information pertaining to the referral comprises both clinical presentation e.g. positive IRT, meconium ileus, and pertinent family history e.g., sibling / parent with CF. The result information specifies the level of testing undertaken (i.e., the number of mutations being screened for), any mutations found, and residual risk calculated.

## Data quality

### Consistency

Much of the information in the CF gene test record was highly heterogeneous, as represented in Table 1. Inconsistent data entry presented a considerable data quality issue. Some of this variation reflects the longitudinal nature of the records, i.e. updates in practice over time, such as adopting different *CFTR* variant naming schemes, but most is likely attributable to inconsistent protocols over the 30-year collection period. Some fields are populated with a variety of synonymous entries that may be interpretable given context (see Table 1, Reason and Result columns) but are not internally consistent enough to serve as reliable annotations.

### Formatting

To facilitate data cleaning, we reformatted the records from an SQL table to a comma separated tabular format with one test per row. The spatial organisation of the initial report dictated the columns created, as visualised in Table 1. PII fields were distributed across separate columns, but free-hand clinical notes were preserved in a small number of large unstructured text fields. These fields encompassed a broad range of information, including clinical context, test result and, problematically, unique identifiers for the patients, their relatives, and in some cases their clinicians too. This particular characteristic resulted in a lack of machine readability and presented a significant challenge to importing the records to the SAIL databank.

## Data source conclusion

Clinical scientists generate reports that are formatted according to the needs of the clinician using them, prioritising human readability and focus on the individual or family. These features contrast with the needs of the data scientist, who requires machine readability, clear separation of variables, and population level focus. Whilst the records may be high quality clinical resources, and there has been increasing standardisation over time, historical records bring pervasive structural issues that limit their quality and utility as a data resource.

# Methods and materials

## Pseudonymisation

The SAIL Databank's pseudonymisation protocol was developed in conjunction with Health Informatics Wales (HWIS), now Digital Health and Care Wales (DHCW) [4]. The procedure generates an anonymised linkage field (ALF) from key unique identifiers, such that an ALF represents a unique individual consistently across linked datasets.

# Obstacles and challenges

## Errant identifiers

A key requirement of SAIL is clear adherence to the separation principle [5, 6]. The noted mixing of patient and family PII with clinical data in large free text columns was a significant obstacle to satisfying this separation principle. An "errant identifier" refers to the presence of unique identifiers in fields containing clinical data.

Table 1: Synthetic representation of the contents, formatting, and structure of the CF test records

| Forename | Surname | Reason | Result | Notes |
|---|---|---|---|---|
| **JOHN** | SMITH | fam_hist | cf_carrier | JOHN’s son, ROBERT SMITH, has cystic fibrosis. *c.1521_1523delCTT* variant has been identified. ROBERT’S paternal aunt, **CAROLINE**, died of cystic fibrosis. |
| **JANE** | SMITH | susp_carrier | 1 | JANE’s son, ROBERT SMITH, has cystic fibrosis. Her partner, JOHN SMITH, is also a carrier for the *c.1521_1523delCTT* variant. JANE is also a carrier. |
| **ROBERT** | SMITH | IRT_pos | IRT 2 mut | ROBERT has presented with [symptoms] and received a positive IRT test. His father, JOHN SMITH, and mother, JANE SMITH, are carriers for *p.(F508del)*. |
| **DAVID** | SMITH | CF | 1_mut | DAVID’s son, JOHN SMITH, is a carrier of DF508 and his daughter **CAROLINE** died of CF. His grandson ROBERT has cystic fibrosis. DAVID is a carrier of *DF508*. His wife **AUDREY** is deceased. |
| **ANDREW** | JONES | fertility | one/N | ANDREW and his partner have been experiencing fertility issues. Familial variant detected. |
| **DELIA** | JONES | FH | NN | DELIA’s cousin, **ALEXIA JONES,** died of CF. Her and her husband **ANDY** want to understand their CF risk. No variants detected. ANDREW is a carrier for the *c.1521_1523delCTT* variant. |
| **MARY** | JONES | family | pos | MARY JONES is the father of ANDREW JONES. She is a carrier for the *c.1521_1523delCTT* variant. |

There are synonymous entries in the Reason and Result fields: “fam_hist”, “FH”, and “family” all indicate testing due to family history. Separately, “one/N”, “1_mut”, and “cf_carrier” indicate a single mutation found. Entries such as “pos” are ambiguous. The variant p.(F508del) is referred to by both its DNA and legacy names, c.1521_1523delCTT and DF508 respectively. The Notes column mixes unique identifiers such as forename and surname with clinical data, such as the variants identified. Bold names are individuals without a test record or who have been referred to by a common common permutations of their forename, and are resultantly not indexed by the Forename/Surname columns. The case of Andrew Jones demonstrates a patients test result being recorded in a relatives report. This table is not intended to be exhaustive. IRT = immunoreactive trypsin; FH = family history; NN = Normal.

It was found that addressing this problem could take two forms: (1) white-listing (positive selection) data to be carried forward into SAIL through specific annotation, or (2) detection and scrubbing (negative selection) of the errant identifiers so that the remaining cleaned data could be taken forward. The latter approach was explored first as it seemed most amenable to automation.

Any iterative regular expression (regex) approach designed to detect and scrub errant identifiers would require a comprehensive reference list to match against, but no sufficiently comprehensive list was available or possible to generate. The list of patients who had received testing was not sufficient because errant identifiers often named individuals not in the testing pool, e.g., untested relatives of the patient (see Table 1 for example names in bold), the attending clinician, etc.

A *de novo* natural language processing (NLP) detection approach was explored inspired by similar work elsewhere [2, 7]. It was initially hypothesised that named entity recognition (NER) coupled with part-of-speech (POS) tagging could identify individuals by proper nouns, and use referential language to indirectly detect names e.g. verbs (Joe Doe *attended*, *presented with* etc.), copulas (where the name is the subject, e.g. Joe *is,* Joe *will be*), and possessives (has, have. However, this approach came with no guarantee to remove all errant identifiers, which was found to be incompatible with the governance requirements of SAIL. Therefore, a more labour-intensive approach of manually positively selecting non-PII data to be imported to SAIL was opted for as this would guarantee de-identification, since PII could be deliberately avoided. Vital fields were defined as genotype, testing level (number of CF variants tested for) and referral category i.e., family screening, newborn screening, etc. and were manually annotated across 16,181 records. From initial contact to completing annotation, this process took approximately 160 human work hours.

## Absent identifiers

Out of 16,181 records, approximately 5,900 lacked an NHS Number, which is used for determinative matching to generate an ALF. A subset of patients without an NHS number were identified in their test results as living outside of the UK, but the vast majority could not be explained. Absence of NHS Number may be ameliorated by comparison of other unique identifiers (e.g. date of birth), but if these were not available, the individual in question could not be linked to other datasets, resulting in a practical exclusion from the project. Prenatal cases were common in this group, given they lack a birth date and often a forename.

In pregnancies where there is a family history of CF or foetal symptoms (such as echogenic bowel), the first step is testing the mother. If the mother is at least a carrier, the foetus may have CF testing through amniocentesis. In these cases, the test record usually lists no forename, no date of birth, and no NHS Number. A similar issue was observed for babies tested as part of newborn screening (NBS): if a name had not been assigned, the child was often recorded as some variant of “Baby-of”, then the mothers name or simply the family name. If the child did not receive a follow-up test at birth to restore this data, then this record would result in a non-match in SAIL and exclusion from the project.

### Cross-indexing

Where couples were tested together, it was common for the results for one individual to be explicitly included in the report for their partner, so that genetic counsellors could approach both as a single case without missing any information. However, this cross-linking of data introduced error to the manual annotation of test result because it disconnected the name the report was filed under from that persons test result (see Table 1, under Andrew Jones). Following anonymisation, any residual error not caught during annotation checking would lead to distortions in the data.

### Mutual exclusivity

Whilst the mutual exclusion of access to identifiable and anonymised data is a clearly justified security policy, in this case it created two issues. Firstly, annotation of the raw data could only proceed in one stage before the data was uploaded to SAIL. Given the project analysis could only be undertaken in SAIL, this limited the amount of time that could be spent on annotation and therefore limited the breadth of information taken forward to the minimum necessary for the project. Secondly, annotations could not be later validated against the identifiable data they were based on, which limited the reliability of the anonymised data.

### CFTR variant nomenclature standardisation

Genetic variants have historically been referred to by the change they induce at the protein level e.g. F508del/DF508/ΔF508, or the cDNA level. These are referred to as ‘legacy’ names and their usage tends to be local unless sufficiently harmonised through common use. The Human Genome Variation Society (HGVS) provides guidelines for consistently naming variants at the cDNA, RNA, genomic, and protein levels e.g. c.1521_1523del (p.Phe508del) [8] and these are the preferred reporting standard.

A key utility point of the CF test record are evaluating change in CF population genetics over the period of testing, but use of legacy or HGVS standard formats was variable through time. To ensure variants could be consistency counted, a standardising resource that could translate across all schemas was required.

A thesaurus was created by linking two discrete *CFTR* variation databases: CFTR2 [9, 10] and a Simple ClinVar extract focusing on pathogenic and likely pathogenic variants [11]. In total, 1108 *CFTR* variants were identified across the two resources, with 417 shared by both databases. DNA name, protein name, and legacy names were not symmetrically available in both. A third resource, the Mutalyzer batch interpreter [12], was used to validate found variants and provide comprehensive entries across naming schemes. Each annotated variant in the CF Test Record was then linked to the thesaurus using each instance of a name for full readability regardless of naming scheme. It was found that even when similar formats were used across reference resources, minor adjustments in syntax were necessary e.g., G>A vs G->A. Harmonising minor details such as this were crucial for efficient linkage of variant names.

## Conclusions and recommendations

A comprehensive data linkage architecture promotes versatile research infrastructure and accurate insights. Whilst new data systems can readily adopt standards and protocols to better anticipate the requirements of anonymised linkage [13], historical data must be carefully adapted to unlock its value, and the complexity of this task should not be underestimated. We have described 30 years worth of Welsh CF gene test records as a novel resource for anonymised data linkage and given an account of the challenges associated with adapting it for this purpose. Future efforts to liberate analogous datasets focused on different conditions will benefit from our recommendations.

Firstly, we have found that dealing with errant identifiers requires positive selection of pertinent data, as negative selection is unlikely to be comprehensive enough to satisfy the separation principle. NLP tools for annotation are unlikely to fully replace manual annotation because of the strict data governance standards of existing linkage frameworks. That said, the development of such tools will be a key part of supporting data adaptation efforts in the future and will be delivered through collaboration between data scientists, clinical scientists, and computer scientists.

Secondly, mutual exclusivity of access to identifiable and anonymised data has the potential to create problematic bottlenecks that may limit the scope of data available for research and the reliability of annotations. Governance officers should be made aware of the drawbacks of this practice and should consider allowing temporary ‘look-back’ access arrangements to enable validation of anonymised resources. Research groups should make provision for more contact time with the identifiable data than is estimated to be required, be clear on what fields are vital, and document as much detail on their encoding practices as possible for post-anonymisation reference.

Thirdly, data controllers should consider linkage as a future destination for their data and adapt their standards accordingly to facilitate it. For historical data, this should take the form of providing documentation such as data dictionaries, metadata on encoding practices through time, data maps, etc. These resources will enhance efforts to adapt historical datasets by helping to assess their quality, to locate and characterise them, and to anticipate resource requirements.

Finally, a challenge specific to genetic test records is standardising the nomenclature of genetic variants. For CF, where there are many causative variants, there was no single comprehensive resource to enable translation of nomenclatures, and this is likely the case for other genetic conditions. We recommend that this issue is anticipated ahead of time and accounted for with the preparation of thesaurus-like resources.

## Acknowledgements

The authors would like to thank Liz Merrifield and Dr Iain Perry for their constructive feedback, and Prof. Peter Christen and Rainer Schnell for their insightful editor comments. We acknowledge the vital support of Sian Morgan at AWMGS, Natalie Richards at the Cardiff Joint Research Office, and Cardiff University for sponsorship of the study. This work was supported by Wales Gene Park which is funded by Health and Care Research Wales.

## Statement on conflicts of interest

The authors have no conflicts to declare.

## Ethics statement

Ethical approval was obtained for researcher access to AWMGS data and to conduct this research project from the Central Bristol Research Ethics Committee (Central Bristol REC 23/SW/0010). Furthermore, section 251 exemption was sought from the Confidentiality Advisory Group (CAG) (23/CAG/0012) as explicit research consent had not been granted from patients directly and could not practicably be sought due to the scale and longitudinal nature of the records. RM is funded by a Health and Care Research Wales PhD Studentship (HS-22-20).

## Data availability statement

Researcher access to personal patient data was organised in collaboration with Cardiff and Vale University Health Board (CAVUHB), NHS Wales R&D, and Cardiff University. Though AWMGS are a pan-Wales service, the Data Controllership lies with CAVUHB. As such, none of the data used in this study are currently available for wider sharing. The archiving of research data in SAIL may present the opportunity for sharing with other groups in future.

## References

1. Griffiths R, Schlüter DK, Akbari A, Cosgriff R, Tucker D, Taylor-Robinson D. Identifying children with Cystic Fibrosis in population-scale routinely collected data in Wales: A Retrospective Review. Int J Popul Data Sci 2020;5. https://doi.org/10.23889/ijpds.v5i1.1346

2. Lacey AS, Fonferko-Shadrach B, Lyons RA, Kerr MP, Ford DV, Rees MI, et al. Obtaining structured clinical data from unstructured data using natural language processing software: IJPDS (2017) Issue 1, Vol 1:359 Proceedings of the IPDLN Conference (August 2016). Int J Popul Data Sci 2017;1. https://doi.org/10.23889/ijpds.v1i1.381

3. Ford DV, Jones KH, Verplancke J-P, Lyons RA, John G, Brown G, et al. The SAIL Databank: building a national architecture for e-health research and evaluation. BMC Health Serv Res 2009;9:157. https://doi.org/10.1186/1472-6963-9-157

4. Lyons RA, Jones KH, John G, Brooks CJ, Verplancke J-P, Ford DV, et al. The SAIL databank: linking multiple health and social care datasets. BMC Med Inform Decis Mak 2009;9:3. https://doi.org/10.1186/1472-6947-9-3

5. Kelman CW, Bass AJ, Holman CDJ. Research use of linked health data–a best practice protocol. Aust N Z J Public Health 2002;26:251–5. https://doi.org/10.1111/j.1467-842x.2002.tb00682.x

6. Christen P, Ranbaduge T, Schnell R. Linking Sensitive Data: Methods and Techniques for Practical Privacy-Preserving Information Sharing: Synopsis by Kerina Jones. Int J Popul Data Sci 2021;6.

7. Fonferko-Shadrach B, Lacey AS, Roberts A, Akbari A, Thompson S, Ford DV, et al. Using natural language processing to extract structured epilepsy data from unstructured clinic letters: development and validation of the ExECT (extraction of epilepsy clinical text) system. BMJ Open 2019;9:e023232. https://doi.org/10.1136/bmjopen-2018-023232

8. den Dunnen JT, Dalgleish R, Maglott DR, Hart RK, Greenblatt MS, McGowan-Jordan J, et al. HGVS Recommendations for the Description of Sequence Variants: 2016 Update. Hum Mutat 2016;37:564–9. https://doi.org/10.1002/humu.22981

9. Castellani C. CFTR2: How will it help care? Paediatr Respir Rev 2013;14:2–5. https://doi.org/10.1016/j.prrv.2013.01.006

10. Sosnay PR, Salinas DB, White TB, Ren CL, Farrell PM, Raraigh KS, et al. Applying Cystic Fibrosis Transmembrane Conductance Regulator Genetics and CFTR2 Data to Facilitate Diagnoses. J Pediatr 2017;181:S27-S32.e1. https://doi.org/10.1016/j.jpeds.2016.09.063

11. Pérez-Palma E, Gramm M, Nürnberg P, May P, Lal D. Simple ClinVar: an interactive web server to explore and retrieve gene and disease variants aggregated in ClinVar database. Nucleic Acids Res 2019;47:W99–105. https://doi.org/10.1093/nar/gkz411

12. Lefter M, Vis JK, Vermaat M, den Dunnen JT, Taschner PEM, Laros JFJ. Mutalyzer 2: next generation HGVS nomenclature checker. Bioinforma Oxf Engl 2021;37:2811–7. https://doi.org/10.1093/bioinformatics/btab051

13. Christen P, Schnell R. Thirty-three myths and misconceptions about population data: from data

capture and processing to linkage. Int J Popul Data Sci 2023;8. https://doi.org/10.23889/ijpds.v8i1.2115

## Abbreviations

| Acronym: | Meaning |
|---|---|
| ALF: | Anonymised linkage field |
| AWMGS: | All-Wales Medical Genomics Service |
| CAG: | Confidentiality Advisory Group |
| CAVUHB: | Cardiff and Vale University Health Board |
| CF: | Cystic Fibrosis |
| CFTR/*CFTR*: | Cystic fibrosis transconductance regulator; non-italics represents protein; italics represents gene |
| DHCW: | Digital Health and Care Wales |
| HGVS: | Human Genome Variation Society |
| HWIS: | Health Wales Informatics Service (now Digital Health and Care Wales) |
| IRT: | Immunoreactive trypsin |
| LIMS: | Laboratory information management system |
| NBS: | Newborn screening |
| NER: | Named entity recognition |
| NHS: | National Health Service |
| NLP: | Natural language processing |
| PII: | Personal identifiable information |
| POS: | Part of speech |
| REC: | Research Ethics Committee |
| SAIL: | Secure Anonymised Information Linkage |
| SQL: | Structured Query Language |