Short report

Algorithmic approach to finding people with multiple sclerosis using routine healthcare data in Wales

Richard Nicholas ,[1] Emma Clare Tallantyre ,[2] James Witts,[3] Ruth Ann Marrie ,[4] Elaine M Craig ,[3] Sarah Knowles,[3] Owen Rhys Pearson ,[5] Katherine Harding,[6] Karim Kreft,[2] J Hawken,[2] Gillian Ingram,[5] Bethan Morgan,[7] Rodden M Middleton ,[3] Neil Robertson,[2] UKMS Register Research Group[8]

[1]Division of Neuroscience, Department of Brain Sciences, Imperial College London, London, UK
[2]Division of Psychological Medicine and Clinical Neurosciences, School of Medicine, Cardiff University, Cardiff, UK
[3]Population Data Science, Singleton Park, Swansea University Medical School, Swansea, UK
[4]Departments of Medicine and Community Health Sciences, University of Manitoba Max Rady College of Medicine, Winnipeg, Manitoba, Canada
[5]Department of Neurology, Swansea Bay University Health Board, Swansea, UK
[6]Royal Gwent Hospital, Aneurin Bevan University Health Board, Newport, UK
[7]Uplands and Mumbles Surgery, Swansea Bay University Health Board, Swansea, UK
[8]UK MS Register, Swansea, UK

Correspondence to
Dr Rodden M Middleton; R.M.Middleton@swansea.ac.uk

Received 30 January 2024
Accepted 29 April 2024
Published Online First 23 May 2024

To cite: Nicholas R, Tallantyre EC, Witts J, *et al. J Neurol Neurosurg Psychiatry* 2024;**95**:1032–1035.

ABSTRACT

Background Identification of multiple sclerosis (MS) cases in routine healthcare data repositories remains challenging. MS can have a protracted diagnostic process and is rarely identified as a primary reason for admission to the hospital. Difficulties in identification are compounded in systems that do not include insurance or payer information concerning drug treatments or non-notifiable disease.

Aim To develop an algorithm to reliably identify MS cases within a national health data bank.

Method Retrospective analysis of the Secure Anonymised Information Linkage (SAIL) databank was used to identify MS cases using a novel algorithm. Sensitivity and specificity were tested using two existing independent MS datasets, one clinically validated and population-based and a second from a self-registered MS national registry.

Results From 4 757 428 records, the algorithm identified 6194 living cases of MS within Wales on 31 December 2020 (prevalence 221.65 (95% CI 216.17 to 227.24) per 100 000). Case-finding sensitivity and specificity were 96.8% and 99.9% for the clinically validated population-based cohort and sensitivity was 96.7% for the self-declared registry population.

Discussion The algorithm successfully identified MS cases within the SAIL databank with high sensitivity and specificity, verified by two independent populations and has important utility in large-scale epidemiological studies of MS.

INTRODUCTION

The challenge of monitoring changing patterns of disease at population levels[1] arises from variations in healthcare systems, how data are collected in community and hospital settings and the need to verify findings through capture-recapture methodology.[2] Anonymised repositories of highly codified 'routine' data provides opportunities for incidence and prevalence monitoring or tracking impacts of pandemics on a population level but adds complexity varying both by the system and by how they are maintained. Repositories exist because of the need for accurate audit, reporting, health surveillance and billing data. Where the reporting of the condition is not mandated, cases can be missed and in public health systems where insurance or reimbursement do not drive reporting, the priority of how treatment codes are applied and reporting can vary.

Ascertainment of cases of multiple sclerosis (MS) is challenging since diagnostic criteria have changed over time and the coding systems for the detail required for accurate disease subtypes are rudimentary. Since 1965, the MS diagnostic approach has evolved from purely clinical criteria where a diagnosis could take years, to criteria that allow integration of paraclinical data into the diagnostic process, including allowing diagnosis after a single clinical event.[3]

Wales is a country, within the UK, with a population of 3.2 million. In common with the UK, it shares a National Health Service (NHS). Uniquely, in Wales, the Secure Anonymised Information Linkage (SAIL)[4] is a data repository from sources including the Welsh NHS (hospitals and specialist services) and from general practice. Data are collected from the Welsh NHS as International Classification of Diseases (ICD) V.10[5] codes, and from more than 85% of general practitioners (GPs) as 'Read' codes,[6] a UK-specific general practice coding system. There are no older coding systems (ICD 8,9) as supplying hospital systems carried out this transition prior to the creation of SAIL. A crucial element of SAIL is that diverse datasets can be robustly linked to existing individual patients. There are several approaches to finding people with MS (pwMS) within routine data repositories. Iterating on Al-Sakran and colleagues' work in Manitoba[7] we developed an algorithm to identify MS cases within Wales. The performance of the algorithm was tested against two other independent MS datasets.

METHODS

Study setting

We used a cross-sectional population-based cohort study to develop an algorithm to identify pwMS living in Wales, using SAIL. SAIL uses a trusted third party (NHS Wales) to implement a unique Anonymous Linking Field (ALF) as records are loaded into the repository.[4] Users of SAIL are not allowed to use any method to identify patients within it. We used the following datasets:

Welsh Demographics Service Dataset

Individuals registered with Welsh GPs (https://data.sail.ukserp.ac.uk/Asset/View/20).

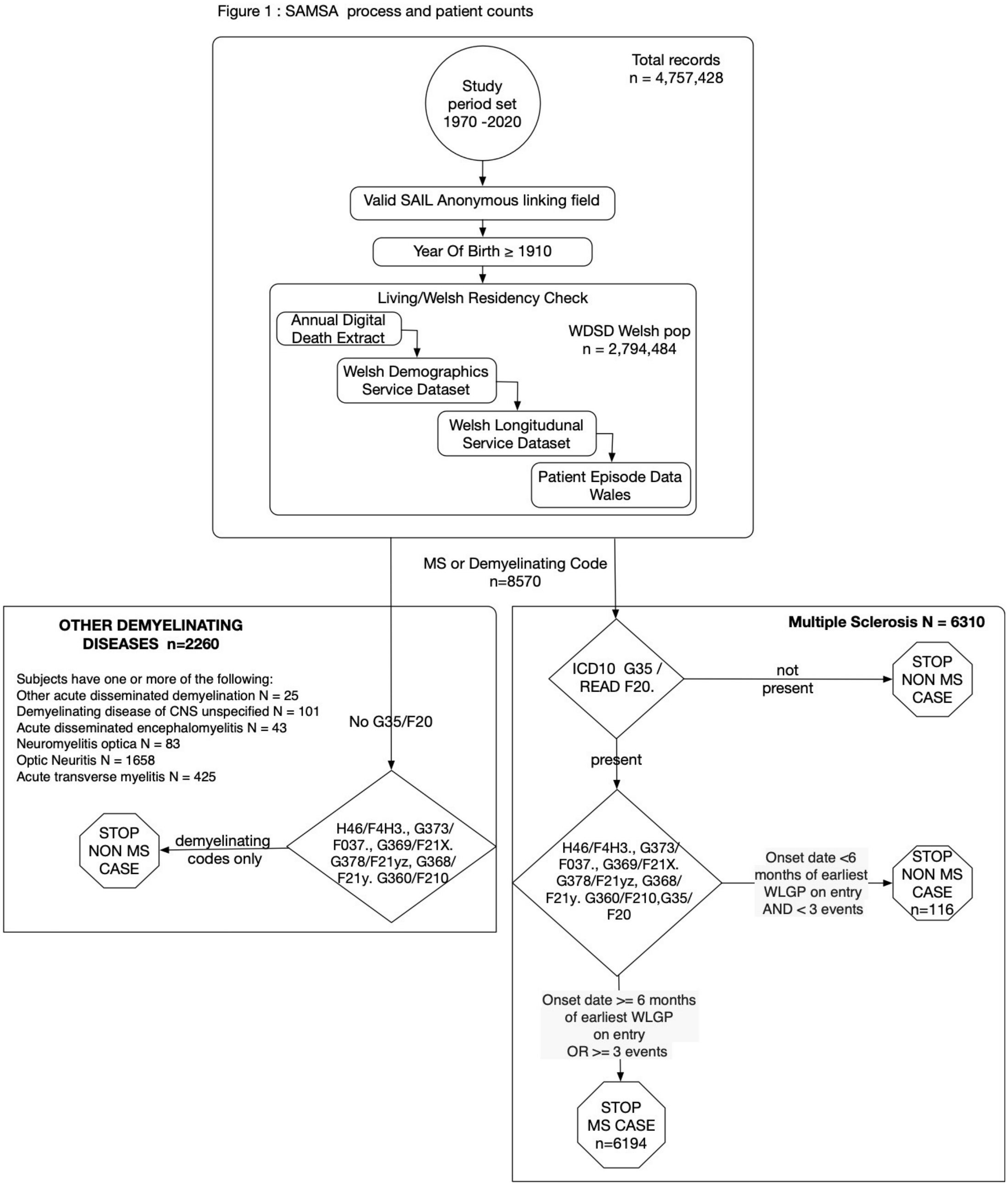
Figure 1 : SAMSA process and patient counts
Study period set 1970 -2020
Total records n = 4,757,428
Valid SAIL Anonymous linking field
Year Of Birth ≥ 1910
Living/Welsh Residency Check
Annual Digital Death Extract
WDSD Welsh pop n = 2,794,484
Welsh Demographics Service Dataset
Welsh Longitudunal Service Dataset
Patient Episode Data Wales
MS or Demyelinating Code n=8570
OTHER DEMYELINATING DISEASES n=2260
Subjects have one or more of the following:
Other acute disseminated demyelination N = 25
Demyelinating disease of CNS unspecified N = 101
Acute disseminated encephalomyelitis N = 43
Neuromyelitis optica N = 83
Optic Neuritis N = 1658
Acute transverse myelitis N = 425
No G35/F20
STOP NON MS CASE
demyelinating codes only
H46/F4H3., G373/ F037., G369/F21X. G378/F21yz, G368/ F21y. G360/F210
Multiple Sclerosis N = 6310
ICD10 G35 / READ F20.
not present
STOP NON MS CASE
present
H46/F4H3., G373/ F037., G369/F21X. G378/F21yz, G368/ F21y. G360/F210,G35/ F20
Onset date <6 months of earliest WLGP on entry AND < 3 events
STOP NON MS CASE n=116
Onset date >= 6 months of earliest WLGP on entry OR >= 3 events
STOP MS CASE n=6194

Figure 1 Flowchart of case finding for each of the algorithms. ICD, International Classification of Diseases; MS, multiple sclerosis; SAIL, Secure Anonymised Information Linkage; WDSD, Welsh Demographics Service Dataset; WLGP, Welsh Longitudinal General Practice Dataset;CNS, Central Nervous System; SAMSA, SAil Multiple Sclerosis Algorithm.

Patient Episode Data for Wales

Routine and emergency hospital admissions and 'spells of care' that include transfers between wards and discharge information (https://data.sail.ukserp.ac.uk/Asset/View/15).

Welsh Longitudinal General Practice Dataset

Welsh Primary Care data covering ~85% of GP practices (https://data.sail.ukserp.ac.uk/Asset/View/17).

Annual Digital Death Extract

UK Deaths (https://data.sail.ukserp.ac.uk/Asset/View/03).

Outpatient Dataset for Wales

Comprises all Welsh Hospitals and includes 'did not attend' information (https://data.sail.ukserp.ac.uk/Asset/View/14).

Independent patient with MS datasets

The South-East Wales MS cohort, comprises clinical data from the Cardiff area in South Wales, updated by clinical and research staff based on patient contact (Ethics: 19/WA/0289, 05/WSE03/111).[8] We included pwMS known to be based in Cardiff diagnosed up to 1 January 2020. Inclusion criteria were appropriate to the epoch of MS diagnosis[9] known to be alive and resident in the Cardiff and Vale area on 31 December 2020.

The UK MS Register patient portal is a nationwide registry capturing data from self-registered pwMS (Ethics: 21/SW/0085). Participants confirm a clinical diagnosis of MS. Data are updated every 6 months. Cases from the UK MS Register (UKMSR) were included if they had a complete date of birth, gender, postcode and Welsh Demographics Service Dataset (WDSD) entry.

Table 1 Sensitivity and specificity of the UK MS Register algorithm in the South Wales and UK MS Register cohorts. The population shown is Cardiff and Vale as of 31 December 2020. Sensitivity=true positive/true positive+false negative. Specificity=true negative/(true negative+false positives), gold standard=South East Wales MS Cohort

	South-East Wales MS cohort		UK MS Register	
Sensitivity of SAIL algorithm	690/713 (96.8%)		808/836 (96.7%)	
Specificity	477 372 / (156+477 372) = 99.9%		–	
	Algorithm (n=690)	**Cohort (n=713)**	**Algorithm (n=808)**	**Cohort (n=836)**
Age (mean±SD), years	54±13.7	53.9±13.8	54.1±11.9	54.1±11.9
Age diagnosis (mean±SD), years	38.2±11.7	37.3±11.4	39±10.7	39.2±10.5

MS, multiple sclerosis; SAIL, Secure Anonymised Information Linkage.

SAIL MS Algorithm

1. SAIL data were included if individuals had a valid ALF code from between January 1970 and December 2020, and a week of birth ≥1910.
2. Next, we checked to see if patients were alive and resident in Wales on 31 December 2020. Datasets were searched in order: Annual Digital Death Extract, WDSD, Welsh Longitudinal General Practice Dataset (WLGP), Patient Episode Data for Wales (PEDW) and checked for dates of death.
3. Next, patients were required to have an ICD-10 code 'G35', 'Multiple Sclerosis' within PEDW/Outpatient Dataset for Wales (OPDW) or a Read code 'F20'. 'Multiple Sclerosis' within WLGP.
4. Once the ICD-10 code G35 was established, the algorithm identified the earliest code from the following list to determine onset date as used previously,[10] optic neuritis (H46/F4H3.), acute transverse myelitis (G373/F037.), acute disseminated encephalomyelitis (G369/F21X.), demyelinating disease of central nervous system not otherwise specified (G378/F21yz), other acute disseminated demyelination unspecified (G368/F21y.), MS (G35/F20.) or neuromyelitis optica (G360/F210).[10]
5. Cases were included if:
   a. There were <3 F20./G35 codes and the established onset date was ≥6 months of the earliest WLGP entry for the patient
   b. OR there were ≥3 F20./G35 codes.

Data analysis

Provisioned data are stored and accessed via the SAIL Secure eResearch Platform.[11] All analyses were conducted using R V.4.1.3. 95% CIs were calculated for prevalence and incidence using Poisson's method. Using the following formulas sensitivity (true positive/(true positive+false negative)) and specificity (true negative/(false positive+true negative)) were calculated.

RESULTS

Incidence and prevalence in an algorithmically identified Welsh MS population

Of the SAIL population of 4 757 428 subjects, the algorithm identified 6194 prevalent cases of MS within Wales on 31 December 2020 (figure 1). Using WDSD for those in Wales in 2020, 209 were incidents (diagnosed in 2020). Given the WLGP population size of 2 794 484 at the end of 2020, incidence is estimated at 7.48 (95% CI 6.5 to 8.56) per 100 000 and prevalence 221.65 (95% CI 216.17 to 227.24) per 100 000.

Comparison of the algorithm versus two independent Welsh MS datasets

We used two validation cohorts to confirm the diagnostic accuracy of the SAIL algorithm (table 1). Of 713 in the South-East Wales MS cohort, 690 (96.8% sensitivity) were identified by the SAIL algorithm. Using the Cardiff dataset as a gold standard 156 people were identified by the algorithm as having MS, but were not in the South East Wales MS dataset, giving a specificity of 99.9%. 69 of the 156 'false positives' had Cardiff hospital data, but were unknown to the MS Service. The remainder were only known to primary care. Among the 836 Welsh pwMS from the UKMSR who had self-registered MS via an online portal, 808 (96.7% sensitivity) were identified by the SAIL algorithm. Specificity analysis was not applied to UKMSR since this is not a population-based cohort. For both populations age of diagnosis was similar between the algorithm and the cohort.

DISCUSSION

We describe an algorithm to identify pwMS within a national routine data repository. We used a clinician-confirmed population-based cohort, and a self-declared previously validated[12] cohort to validate the algorithm, confirming high sensitivity and specificity. We were able to demonstrate that the age at diagnosis calculated by the algorithm is similar to the age at diagnosis self-reported by patients and confirmed by clinicians within both validation cohorts.

Population-wide repositories of health data provide a valuable opportunity to study trends in disease patterns over time. However, identification of all subjects with a disease depends on the reach of the system and how regularly and rigorously it is maintained. A claims-based registry in Canada has developed a reliable methodology for capturing incidents and prevalent cases of MS.[10] However, its transferability is affected by different coding mechanisms and collection drivers. SAIL contains primary and secondary care health data[4] but not disease-modifying therapy prescribing/billing data unlike Canada's insurance-based healthcare system, where it can be used as a confirmatory code for MS diagnosis. In contrast, in SAIL diagnostic confirmation is based, potentially more reliably, on healthcare professional input. Given these differences, we were able to develop an algorithm within this UK-based public health system, relying on a combination of hospital and GP reporting that was able to capture>96% of known MS cases.

Using the SAIL algorithm, we estimated the incidence of MS in our region to be 7.48 (95% CI 6.5 to 8.56) per 100 000. An earlier SAIL study using only hospital-identified MS cases and only one MS code found that Welsh incidence was 9.10 (95% CI 8.80 to 9.40) per 100 000.[13] Whereas a population-based study undertaken in South Wales in 2007 found the incidence was 9.65 (95% CI 7.71 to 13.1) per 100 000.[14] Our lower incidence reflects the more conservative approach taken to case ascertainment in this algorithm, but we have confirmed its sensitivity and specificity in two independent cohorts. Our prevalence finding of 221.65 per 100 000 people for Wales is higher than other reported figures for the UK as a whole at 199 per 100 000,[15] but consistent with the predictions made in the South-East Wales region,[14] suggesting a rise of MS prevalence to 260 per 100 000 population by 2028–2048.

We also identified a group of 2260 people in Wales who had a demyelination code but did not have MS. By confirming the high sensitivity and specificity of our algorithm versus a clinically diagnosed MS population, we conclude that we were correct to exclude this group at this point, but it will inevitably contain people who later go on to develop MS.

In this study, we present a robust, sensitive algorithm to ascertain cases of MS in large populations that, with pragmatic adaptations, could be adapted to effectively identify cases in other large geographical areas with similarly structured data systems.

Acknowledgements This study makes use of anonymised data held in the Secure Anonymised Information Linkage (SAIL) Databank. We would like to acknowledge all the data providers who make anonymised data available for research. All interpretations of SAIL data are the authors' own.

Collaborators Professor Nikos Evangelou, Kellie Allen, Professor Klaus Schmierer, Dr Ian Galea, Dr Matt Craner, Professor Jeremy Chataway, Dr Gavin McDonnell, Ms Annemieke Fox, Dr Heather Wilson, Dr David Rog, Dr Chris Kipps, Dr Andrew Gale, Dr Monica Marta, Sarah Fuller, Dr Judy Archer, Dr Brendan McLean, Dr Agne Straukiene, Dr Joe Guadango, Jo Kitley, Dr Andrew Graham, Dr Carlo Canepa, Dr Helen Ford, Professor Alasdair Coles, Professor H Emsley, Professor Jeremy Hobart, Julie Foxton, Dr Dreedharan Harikrishnan, Dr Laura Petzold, Dr Tim Harrower, Professor Ruth Dobson London, Dr Zbigniew Slowinski, Professor Basil Sharrack.

Contributors RN, RMM, RAM, ECT and NR: Conceptualisation, Resources, Writing, Review and editing. GI, ORP, KH, JH, KLK, BM: Review and editing. JW, SK, EMC: Data Analysis, Review and editing.

Funding This study was funded by Multiple Sclerosis Society (147).

Competing interests None declared.

Patient consent for publication Not applicable.

Provenance and peer review Not commissioned; externally peer reviewed.

ORCID iDs
Richard Nicholas http://orcid.org/0000-0003-0414-1225
Emma Clare Tallantyre http://orcid.org/0000-0002-3760-6634
Ruth Ann Marrie http://orcid.org/0000-0002-1855-5595
Elaine M Craig http://orcid.org/0000-0002-3432-9942
Owen Rhys Pearson http://orcid.org/0000-0002-2712-0200
Rodden M Middleton http://orcid.org/0000-0002-2130-4420

REFERENCES

1 Pericleous M, Kelly C, Odin JA, *et al*. Clinical Ontologies improve case finding of primary biliary cholangitis in UK primary and secondary care. *Dig Dis Sci* 2020;65:3143–58.
2 Hook EB, Regal RR. Capture-recapture methods in epidemiology: methods and limitations. *Epidemiol Rev* 1995;17:243–64.
3 Polman CH, Reingold SC, Banwell B, *et al*. Diagnostic criteria for multiple sclerosis: 2010 revisions to the McDonald criteria. *Ann Neurol* 2011;69:292–302.
4 Jones KH, Ford DV, Jones C, *et al*. A case study of the secure anonymous information linkage (SAIL) gateway: A privacy-protecting remote access system for health-related research and evaluation. *J Biomed Inform* 2014;50:196–204.
5 WHO. ICD-10 Version:2010. ICD-10 Version2010, 2010. Available: https://icd.who.int/browse10/2010/en [Accessed 28 Apr 2021].
6 Nhs NB. What are the read codes *Health Libraries Review* 1994;11:177–82.
7 Al-Sakran LH, Marrie RA, Blackburn DF, *et al*. Establishing the incidence and prevalence of multiple sclerosis in Saskatchewan. *Can J Neurol Sci* 2018;45:295–303.
8 Harding KE, Ingram G, Tallantyre EC, *et al*. Contemporary study of multiple sclerosis disability in South East Wales. *J Neurol Neurosurg Psychiatry* 2023;94:272–9.
9 Thompson AJ, Banwell BL, Barkhof F, *et al*. Diagnosis of multiple sclerosis: 2017 revisions of the McDonald criteria. *Lancet Neurol* 2018;17:162–73.
10 Marrie RA, Yu N, Blanchard J, *et al*. The rising prevalence and changing age distribution of multiple sclerosis in Manitoba. *Neurology* 2010;74:465–71.
11 Jones KH, Ford DV, Ellwood-Thompson S, *et al*. The UK secure eResearch platform for public health research: a case study. *The Lancet* 2016;388:S62.
12 Middleton RM, Rodgers WJ, Chataway J, *et al*. Validating the portal population of the United Kingdom multiple sclerosis register. *Mult Scler Relat Disord* 2018;24:3–10.
13 Balbuena LD, Middleton RM, Tuite-Dalton K, *et al*. Sunshine, sea, and season of birth: MS incidence in Wales. *PLOS ONE* 2016;11:e0155181.
14 Hirst C, Ingram G, Pickersgill T, *et al*. Increasing prevalence and incidence of multiple sclerosis in South East Wales. *J Neurol Neurosurg Psychiatry* 2009;80:386–91.
15 Number of people with MS | Atlas of MS, Available: https://www.atlasofms.org/map/global/epidemiology/number-of-people-with-ms [Accessed 15 Mar 2024].