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

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

Corresponding Author: Rod Middleton, Population Data Science, Swansea University Medical School, Singleton Park, Swansea, UK, SA28PP

Email: r.m.middleton@swansea.ac.uk

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, Swansea University Medical School, Singleton Park, Swansea, UK
4. Departments of Medicine and Community Health Sciences, University of Manitoba Max Rady College of Medicine, Winnipeg, Canada
5. Department of Neurology, Swansea Bay University Health Board, Swansea, UK
6. Aneurin Bevan University Health Board, Royal Gwent Hospital, Newport, UK
7. Swansea Bay University Health Board, Uplands & Mumbles Surgery, Swansea, UK
8. UK MS Register Research Group – see Acknowledgements section

Abstract

Background: Identification of multiple sclerosis (MS) cases in routine health care data repositories remains challenging. MS can have a protracted diagnostic process and is rarely identified as a primary reason for admission to 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 6,194 living cases of MS within Wales on 31st December 2020 (prevalence 221.65 [95%CI 216.17–227.24] per 100,000). Case finding sensitivity and specificity was 96.8% and 99.9% for the clinically validated population-based cohort and sensitivity was 96.7% from the self-declared registry population.

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

Significance of this study

What is already known about this subject:

- Case finding is essential for incidence and prevalence studies with the development of national databases with the power to link to multiple datasources and investigate early disease risk factors additionally it is difficult to identify conditions that have a complex diagnostic process and do not need to be formally reported
- Since 1965, the Multiple Sclerosis 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.
- Case finding of MS in routinely linked data sets has been reliably carried out but is dependent on the coding systems used and types of data available within them.

What are the new findings:

- Validation of the use of routinely linked healthcare data to reliably discover patients with multiple sclerosis at a population level
- Improved estimates of the incidence and prevalence of MS for a nation based on a population level dataset.
- Ability to link disparate datasets from gold standard clinical databases and participant supplied patient reported outcome data to comprehensively validate the algorithm.

How might it impact on clinical practice in the foreseeable future?

- The study presents 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.

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 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 United Kingdom (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) version 10 [5] codes, and from more than 85% of General Practitioners (GP) 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 Marrie 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 utilised 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 utilised the following datasets:

Welsh Demographics Service Dataset (WDSD): Individuals registered with Welsh GPs [https://data.sail.ukserp.ac.uk/Asset/View/20].

Patient Episode Data for Wales (PEDW): 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 (WLGP): Welsh Primary Care data covering ~85% of GP practices. [https://data.sail.ukserp.ac.uk/Asset/View/17]

Annual Digital Death extract (ADDE):UK Deaths [https://data.sail.ukserp.ac.uk/Asset/View/03]

Outpatient Dataset for Wales (OPDW): comprises all Welsh Hospitals and includes 'did not attend' information. [https://data.sail.ukserp.ac.uk/Asset/View/14]

Independent MS patient datasets

The South-East Wales MS cohort, comprises clinical data from 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 1st January 2020. Inclusion criteria were appropriate to epoch of MS diagnosis [9] known to be alive and resident in the Cardiff and Vale area on 31st December 2020.

The UK MS Register (UKMSR) 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 six months. Cases from the UKMSR were included if they had a complete date of birth, gender, postcode and WDSD entry.

SAIL MS Algorithm (Figure 1)

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 31st December 2020 Datasets were searched in order: ADDE, WDSD, WLGP, PEDW and checked for dates of death.

3. Next, patients were required to have an ICD-10 code 'G35', 'Multiple Sclerosis' within PEDW/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. Ninety-five per cent confidence intervals (95%CI) 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.

Figure 1. Flowchart of case finding for each of the algorithms.

Results

Incidence and prevalence in an algorithmically identified Welsh MS population

Of the SAIL population of 4,757,428 subjects the algorithm identified 6,194 prevalent cases of MS within Wales on 31st December 2020 (Figure 1). Using WDSD for those in Wales in 2020, 209 were incident (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 6.5–8.56] per 100,000 and prevalence 221.65 [95%CI 216.17–227.24] 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.

Table 1. Sensitivity and specificity of the UK MS Register algorithm in the South Wales and UK MS Register cohorts. Population shown is Cardiff and Vale as of 31/12/2020. 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	477372 / (156 + 477372) = 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

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 age at diagnosis calculated by the algorithm is similar to 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 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 incident 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 incidence of MS in our region to be 7.48 [95%CI 6.5–8.56] 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 [8.80–9.40] per 100,000 [13]. Whereas a population-based study undertaken in South Wales in 2007 found the incidence was 9.65 [7.71-13.1] 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 2,260 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.

Availability

The R code for the case finding algorithm is available from :
https://github.com/roddenm/CaseFindingMS.

SAIL data are available upon application to the SAIL Information Governance Review Panel. Further details can be found at https://saildatabank.com/application-process/

Disclosures

Contributorship

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. https://icd.who.int/browse10/2010/en (accessed 28 April 2021)

6   Nhs NB. What are the Read Codes? *Health Libr Rev*. 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 J Can Sci Neurol*. 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. 2010;10.

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*. 2008;80:386–91.

15 Number of people with MS | Atlas of MS. https://www.atlasofms.org/map/global/epidemiology/number-of-people-with-ms (accessed 15 March 2024)