OPCRD ADEPT application Study Protocol – V2, 11th Nov 2022

Study Title

A matched case-control study to identify features associated with a group of rare diseases and examine the case identification accuracy of MendelScan, a rare disease case finding tool.

Lay Summary

Rare diseases are individually uncommon, affecting less than 1 person in 2000, however with more than 6,000 diseases they are collectively common. A feature shared by many rare diseases is a long path to diagnosis, typically measured in years or even decades. During this 'diagnostic odyssey' patients experience the many challenges of not having an accurate diagnosis; repeated investigations and referrals; a lack of explanation for their problems, and a lack of expert care and/or treatments. Further, until a diagnosis is made, affected individuals cannot benefit from the support of patient advocacy groups.

MendelScan is a rare disease case finding tool that uses patients' GP records to identify patterns that suggest they may have an undiagnosed rare disease. Identified patient records are then reviewed and a targeted report returned to their GP for suggested next steps including further. In this study we will use patients' GP records in the large primary care research database, OPCRD, to examine the performance of MendelScan for a range of diseases and use the database to support the development of other rare disease detection tools.

Technical Summary

Background

Rare diseases are individually rare but collectively common, with many patients experiencing a substantial delay in diagnosis. Data collected in the electronic health records (such as seeking medical attention for symptoms) has been identified as a source of information that could help expedite the diagnosis of these patients. MendelScan is a rare disease case finding tool that uses routinely collected structured primary care data to flag patients at risk of having a rare disease.

Objectives

Validation of pre-developed MendelScan rare disease algorithms, in a primary care dataset and assess the scope for improving on these algorithms by use of alternative modelling techniques.

Study design

We seek to extract a sub-population of the available OPCRD cohort to perform a series of case-control studies to evaluate and improve upon the performance of MendelScan case-finding algorithms. This is to estimate the potential 'real world' performance of such predefined criteria-based tools if deployed in their target use case, i.e. embedded within general practice software.

These case-control studies will be performed for each rare condition of interest, with a standardised methodology applied. This will include descriptive statistics of cases and controls and assessment of the sensitivity, specificity, and PPV of MendelScan algorithms. In diseases with low prevalence, clinical validation will also involve case review by primary care and disease specific experts.

Setting/Participants

UK General practice, the OPCRD research data set with linked HES data for certain disease cohorts if available.

Participants: Individuals identified by the appropriate SNOMED/Read diagnostic codes for each rare disease matched to controls who were under the care of this practice at a similar time. .

Objectives

Primary objective:
To assess the accuracy for the purpose of validation of MendelScan RD case finding algorithms (see Appendix 1).

Study Background

Rare diseases (RD), defined in the UK and EU as affecting fewer than 1 in 2000 persons, are individually rare but collectively common [1]. With 6,000-8000 RD they affect 3.5–5.9% of the population or 263–446 million persons globally [2]. A common feature across many RD is diagnostic delay, sometimes termed the 'diagnostic odyssey', with cohorts of patients in the UK and US reporting an average 5.6 and 7.6 years delay respectively, and patients typically visiting eight physicians (four primary care and four hospital specialists) and receiving two or three misdiagnoses in advance of their correct diagnosis[3]. Similarly, an EU survey reported that 40% of patients with RD were initially incorrectly diagnosed, and a quarter experienced a diagnostic odyssey of more than 5 years [4]. Early diagnosis is central to achieving better patient outcomes for RD patients, and supporting clinicians (especially those in primary care) to identify unusual patterns and revisit diagnoses is crucial to reducing the 'diagnostic odyssey' [5]. This is a widely recognised challenge, and is the first of four key priorities in the UK Rare Diseases Framework, published in January 2021, with data and digital technologies highlighted as a potential solution [6].

Mendelian is a health technology company and provider of MendelScan, a software platform and Class 1 Medical Device, that runs disease suspicion criteria to detect patients with undiagnosed rare diseases based on their electronic health record (EHR).

Mendelian has developed and digitised suspicion criteria for 25 rare diseases (using SNOMED Clinical Terms), which we seek to robustly validate in UK primary care EHRs. These criteria are currently being deployed and evaluated prospectively, for their 'real-world' clinical validity in a UK primary care pilot. These tools are currently rules-based heuristics, and there may be benefits for some algorithms on applying more flexible methodologies.

Study Design

We seek to use a standardised statistical approach to examine the validity and potential utility of a series of predefined MendelScan tools. These tools are based on set criteria, such as symptoms and age, and are based on clinical expert input and literature review.

For each condition of interest (see **Appendix 1**), we will conduct a retrospective analysis using a case control design. From the entire available OPCRD database, we will identify all confirmed cases of the listed rare diseases with at least 3 years of primary care EHR history prior to the diagnostic date (index date). Confirmed cases are defined by the presence of an appropriate SNOMED/READ code. Each case will be matched to 100 controls registered at the same practice during the same calendar year. Age and sex will not be used for matching as these often feature in the disease algorithms. As we seek to explore the potential 'real world' validity of these tools, which ideally will be embedded in general practice software and run over all patient records, this approach to validation aligns with intended use. As we are not seeking to estimate an odds ratio for a specific parameter/factor, and are not seeking to establish a causal coefficient (rather, assess the sensitivity, etc. of pre-defined case-finding rules), matching on the basis of age and sex and/or adjusting for them is not necessary.

A sensitivity analysis will be performed for a subset of diseases that have a disease specific discriminatory code in ICD-10, where cases can be further defined from HES data.

Study population/Selection of Controls

UK Primary care OPCRD population. All cases with SNOMED/READ diagnostic codes for the rare

diseases (see table) with at least 3 years of EHR before diagnostic date (index date), will be matched to a control population at a ratio of 100 controls to each case. Controls will be practice matched but age and gender will not be used for matching as these feature in the algorithms. Like the cases each control will have at least 3 years of EHR in advance of the index date (date of diagnosis of matched case).

Data/ Statistical analysis

Descriptive statistics for each disease will be performed for both cases and control groups, reporting number (%), mean (SD) and median (IQR) for categorical, normal continuous and non-normal continuous variables, respectively. Missing values will also be presented. Appropriate statistics tests

such as $\chi 2$, t-tests, and analysis of variance tests (ANOVA) will be used to assess the differences between the groups of interest.

Each algorithm will be tested in a population of all controls and known cases of that disease. Analysis will involve construction of 2x2 contingency tables for each disease. Fisher's exact method or Chi^2 method will be performed depending on cell count.

The performance of each MendelScan algorithm will be measured by sensitivity and specificity. As these diseases are rare the primary metric of interest to indicate the clinical value of these algorithms is the positive predictive value, PPV. As the analysis is a case-control study and therefore the prevalence is manipulated by sampling, the PPV and its standard error will be calculated using a standard formula correcting for the sampling fraction of the control population [7].

We will also perform sensitivity analyses cognisant of the possibility that the control populations may contain undiagnosed patients. Therefore, as not all of the control population flagged by the algorithms will be correctly identified as false positives, we will estimate the number of undiagnosed patients for each rare disease in the population, by subtracting the number of diagnosed cases from the expected number for a range of published prevalence figures to give an estimated number of undiagnosed patients.

We will use the sensitivity of the algorithm to identify known cases to give an indication of the proportion of undiagnosed cases the algorithm will identify. We will use these estimates to reassign those ‘false positive’ cases to ‘true positive’ cases, this number will then be added to the known number of cases flagged by the algorithm to calculate the PPV. We will calculate the range of PPV for

the range of population prevalence for the disease and the disease 95% CI for the algorithm sensitivity in known

cases. For some diseases, further algorithm development will be explored as below (**Amendment Statement** and **Amendment to analysis plan**).

Considerations for missing data

For categorical demographic variables such as ethnicity, alcohol or smoking status, an "unknown" category will be coded.

For categorical clinical variables (such as depression, rheumatoid arthritis) which are missing or not recorded, we will assume that not being recorded indicates their absence. This will preserve the sample size and is a common assumption in analyses of large general practice-based EHR research. For continuous variables, including missing variables such as BMI, pulse etc., descriptive statistics will describe the proportion with missing values.

On an individual MendelScan tool basis, if the extent of missing data relevant to the algorithm is extensive, we may consider the use of multiple imputation to replace missing values (e.g. for BMI) and report the ranges of PPV obtained using this approach. Briefly, the PPV and standard error would be estimated in each imputed dataset and combined using Rubin's rules to form a pooled estimate.

Patient group involvement

We have had extensive rare disease patient group involvement, including specific patient insights into early clinical features for diseases including, Fibrodysplasia Ossificans Progressiva (FOP), neuroendocrine tumours (NET), Addison's disease, Niemann Pick Type C. A number of patient group have acted as collaborators on projects and findings will be shared with the patient communities through established links.

Plans for disseminating and communicating study results

Study outcomes will be shared at academic conferences, in peer reviewed journals and with the rare disease patient community through established rare disease patient advocacy links. They will also act as a foundation for the deployment of MendelScan in the NHS, through established collaborations including NHSEI funded projects with Central and South Genetic Medicine Service Alliance (GMSA)

& North East and Yorkshire GMSA.

Limitations of this study design data sources and analytic methods

The primary limitation of the study is its retrospective design, and as a secondary analysis, the completeness of data ascertainment and quality of recording and coding are both restricted and outside our control.

An acknowledged limitation of using general practice databases is the misclassification of diagnosed cases due to non-recording or the use of the wrong or non-specific diagnostic code. In the case of rare conditions this is less of a risk, the presence of a diagnostic code is likely to be accurate. In this study the risk is further mitigated by the process of disease selection for algorithm development taking into consideration the presence of suitable coding for appropriate case ascertainment before disease selection. Furthermore we intend to explore this by performing sensitivity analyses for certain diseases

where there is suitable HES data diagnostic code, to have greater confidence in the accuracy of the primary care diagnostic coding.

The low prevalence of rare diseases also means that very few cases are likely to sit within the control group. An additional limitation of taking an epidemiological approach, the use of 'risk factors' for later disease development, is that a substantial proportion of rare diseases are congenital disease, with patients "affected" at varying degrees from birth.

In terms of the quality of the initial data capture, GP databases are heavily dependent on GP judgement and patient understanding, especially for self-reported variables such as smoking and family history. In terms of completeness, large GP databases have an unavoidable amount of missing data. There is a possibility that the codes entered will be non-specific with/without free text entry and/or be incomplete. In addition, variables such as blood pressure may not be measured consistently for all patients, and may not remain static over several years of follow up.

As detailed in the statistical analysis section, an assessment of algorithm performance is dependent on an assumption of an algorithm's accuracy to identify undiagnosed patients.

This is built on the following assumptions:

- Firstly, an algorithm's sensitivity based on its ability to identify in advance of the diagnostic date of cases is applicable to the undiagnosed population.

- Secondly, the number of undiagnosed cases in a population can be estimated with sufficient accuracy given the uncertainty in published prevalence figures for rare diseases, and their applicability to the study population.

We will reflect this uncertainty by calculations across the 95% CI range for the algorithm sensitivity and the range of expected cases in the population given the prevalence figures for each disease.

Amendment statement

We propose to use the database to support the development of other rare disease detection tools - we are now in a position to focus on this area of the project more closely.

We have found that existing MendelScan tools are limited to using the core clinical features for detection of each rare disease. These core features are defined in the literature.

Our exploratory analysis during the process of this project has identified features which commonly occur in rare diseases but are not being picked up by our existing algorithms. It has become apparent that we are not making full use of the rich data provided in OPCRD and we therefore aim to move towards a data-driven approach to detecting rare diseases rather than relying on rules-based heuristics informed by previous publications.

We believe more advanced methods are required to build on the work we have done with rules based algorithms. We propose to use OCPRD data to develop and validate new modelling tools to classify rare disease patients, the outcome of which will be of high clinical value and provide indications as to appropriate next steps for clinicians (e.g. genetic testing for specific disease, closer monitoring and support).

Finally, we aim to focus on identifying clinical features which have high predictive value and will serve as early predictive markers of rare diseases. Our exploratory work identified features such as depression and anxiety which are commonly present. These features will be quantified by way of incidence and prevalence in rare disease patients compared to control patients in order to assess whether our algorithms can be improved by inclusion of such features.

Updated Lay Summary

Rare diseases are individually uncommon, affecting less than 1 person in 2000, however with more than 7000 diseases they are collectively common. A feature shared by many rare diseases is a long path to diagnosis, typically measured in years or even decades. During this 'diagnostic odyssey' patients experience the many challenges of not having an accurate diagnosis; repeated investigations and referrals; a lack of explanation for their problems, and a lack of expert care and/or treatments. Further, until a diagnosis is made, affected individuals cannot benefit from the support of patient advocacy groups and often experience poor mental health. MendelScan is a rare disease case finding tool that uses patients' GP records to identify patterns that suggest they have an undiagnosed rare disease. Identified patient records are then reviewed and a targeted report returned to their GP for suggested next steps including referrals to specialist services. In this study we will use patients' GP records in the large primary care research database, OPCRD, to examine the performance of MendelScan for a range of diseases. Furthermore, we will use the richness of the OPCRD database to build "clinical profiles" which will capture various elements of a patient's health record (diagnoses, symptoms, healthcare utilisation) and develop more advanced detection tools (including Supervised Machine Learning models). Finally, we aim to identify and assess the predictive ability of clinical features which occur early and frequently in the patient's healthcare journey, but may not be part of

the core features of any specific rare disease (e.g. depression).

Amendment to analysis plan

Case/control selection - scientific judgement will be exercised in selecting control samples appropriate for the statistical approach employed in the project, but may include matching cases and controls on variables such as age, gender, practice. Steps such as out-of-sample validation may include random selection of controls from distinct general practices. In all cases, we will use approaches outlined above to a) ensure undiagnosed rare disease patients are not included in the control sample and b) account for the possibility of undiagnosed rare disease patients being included in the control sample (e.g. using prevalence estimations).

Supervised Machine Learning models that we intend to test the performance of for this purpose may include logistic regression, decision trees and random forests. Evaluation metrics will be extended to include precision (PPV), recall, F1 and F-beta scores, as well as accuracy, sensitivity and specificity.

Clinical features will be created by inclusion of all relevant SNOMED codes. For example a clinical feature of "abdominal pain" would be present if any SNOMED code pertaining to pain in the abdomen

is present, excluding in pregnancy or other circumstances which are deemed clinically distinct. We will work closely with our clinical team to determine the most appropriate selection of codes for each clinical feature.

References

1. Somanadhan S, Nicholson E, Dorris E, Brinkley A, Kennan A, Treacy E, Atif A, Ennis S, McGrath V, Mitchell D, O'Sullivan G, Power J, Lawlor A, Harkin P, Lynch SA, Watt P, Daly A, Donnelly S, Kroll T. Rare Disease Research Partnership (RAinDRoP): a collaborative approach to identify research priorities for rare diseases in Ireland. HRB Open Res. 2020 Nov 11;3:13. doi: 10.12688/hrbopenres.13017.2. PMID: 33299965; PMCID: PMC7702160.

2. Nguengang Wakap S, Lambert DM, Olry A, Rodwell C, Gueydan C, Lanneau V, Murphy D, Le Cam Y, Rath A. Estimating cumulative point prevalence of rare diseases: analysis of the Orphanet database. Eur J Hum Genet. 2020 Feb;28(2):165-173. doi: 10.1038/s41431-019-0508-0. Epub 2019 Sep 16. PMID: 31527858; PMCID: PMC6974615.

3. Shire Rare disease impact report: insights from patients and the medical community. *J Rare Disord.* 2014:1–34. [Google Scholar]

4. Eurordis.org. 2021 [cited 2021 May 25]. Available from: https://www.eurordis.org/IMG/pdf/Fact_Sheet_Eurordiscare2.pdf

5. Evans WR, Rafi I. Rare diseases in general practice: recognising the zebras among the horses. Br J Gen Pract. 2016 Nov;66(652):550-551. doi: 10.3399/bjgp16X687625. PMID: 27789486; PMCID: PMC5072891.

6. Assets.publishing.service.gov.uk. 2021 [cited 2021 May 26]. Available from: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/950651/the-UK-rare-diseases-framework.pdf

7. van Zaane, B., Vergouwe, Y., Donders, A.R.T. *et al.* Comparison of approaches to estimate confidence intervals of post-test probabilities of diagnostic test results in a nested case-control study. *BMC Med Res Methodol* 12, 166 (2012). https://doi.org/10.1186/1471-2288-12-166

Appendix 1

Diseases for analysis

Alpha-1-antitrypsin deficiency

Churg-Strauss syndrome

Common variable immunodeficiency

DiGeorge syndrome (22q11 deletion)

Fibrodysplasia ossificans progressiva

Good syndrome

Hereditary angioedema

Wiskott aldrich syndrome

X-linked agammaglobulinemia

Behcets disease

Dermatomyositis

Loeys-Dietz syndrome

Marfan syndrome

Prader Willi syndrome

Scleroderma

Tuberous sclerosis

Turner syndrome

Addison disease

Ehlers-Danlos syndrome

Hereditary hemorrhagic telangiectasia

Lynch syndrome

Narcolepsy

Neuro endocrine tumors - Midgut - Functional

Niemann-Pick C disease

Primary biliary cirrhosis

Wilson disease

Osteogenesis imperfecta

22Q13, Phelan-Mcdermid Syndrome

Aarschot-Scott Syndrome

ADNP Syndrome

Adult Onset Still's disease

Alagille Syndrome

Alkaptonuria

Alport syndrome

Amyloidosis

Angelman syndrome

Ankylosing spondylitis

Antiphospholipid syndrome

Arnold Chiari Syndrome type 1

Atypical hemolytic uremic syndrome

Axenfeld-Rieger Syndrome

Bardet-Biedl syndrome

Beckwith-Wiedemann syndrome

Celiac disease

Charcot-Marie-Tooth disease

Chronic progressive ophthalmoplegia

Chronic recurrent multifocal osteomyelitis

Cloves Syndrome

CREST syndrome

Crohn's disease / Ulcerative colitis (IBD)

Cushing syndrome

Muscular dystrophies (Incl. Duchenne MD, Becker MD, Limb girdle MD etc)

Dystrophinopathies

Eosinophilic esophagitis

Epidermolysis Bullosa

Fabry disease

Familial hypercholesterolemia

Floating Harbor Syndrome

Focal dystonia

Fragile X syndrome

Friedreich ataxia

Gaucher disease

Giant cell arteritis

Glycogen storage disease type 5 (Mcardle)

Homocystinuria due to cystathionine beta-synthase deficiency

Huntington disease
Hypermobile Ehlers-Danlos Syndrome
Hypoparathyroidism
Hypophosphatasia
Idiopathic pulmonary hemosiderosis
Idiopathic Pulmonary Hypertension
Juvenile arthritis, idiopathic
Kartagener syndrome
Klippel Feil Syndrome
Leber Congenital Amaurosis
Li-Fraumeni syndrome
Long COVID syndrome
Lysosomal acid lipase deficiency
Mccune-Albright Syndrome (Polyostotic
Fibrous Dysplasia)
Metachromatic leukodystrophy
Microscopic polyangiitis
Mitochondrial diseases (Overall)
Moyamoya
Mucopolysaccharidosis (Overall)
Multiple myeloma
Myasthenia gravis
Neurofibromatosis type 1
Neurofibromatosis type 2
Neuromyelitis optica (NMO)
Noonan syndrome
Osteogenesis imperfecta
Paroxysmal nocturnal hemoglobinuria (PNH)
Peutz-Jeghers syndrome
Polymyositis
Pompe Disease
Porphyria, acute intermittent
Primary ciliary dyskinesia
Primary immunodeficiencies (Overall)
SAPHO syndrome

Sclerosing cholangitis

13

Sjogren's Syndrome
Stickler syndrome
Sturge-Weber Syndrome
Takayasu Arteritis
Tarlov Cyst Disease
Temporal arteritis
Tethered Cord Syndrome
Transthyretin-related amyloidosis (TTRA)
Von Hippel Lindau Disease
Waldenstrom macroglobulinemia
Wegener granulomatosis
Whipple disease
William syndrome
X- linked hypophosphatemia
Xeroderma pigmentosum
Zollinger-Ellison syndrome

14