Automated extraction of clinical traits of multiple sclerosis in electronic medical records

Mary F Davis,[1] Subramaniam Sriram,[2,3] William S Bush,[1,4] Joshua C Denny,[4] Jonathan L Haines[1,2]

▶ Additional material is published online only. To view please visit the journal online (http://dx.doi.org/10.1136/amiajnl-2013-001999).

[1]Center for Human Genetics Research, Vanderbilt University Medical Center, Nashville, Tennessee, USA
[2]Department of Neurology, Vanderbilt University Medical Center, Nashville, Tennessee, USA
[3]Vanderbilt Multiple Sclerosis Center, Vanderbilt University Medical Center, Nashville, Tennessee, USA
[4]Department of Biomedical Informatics, Vanderbilt University Medical Center, Nashville, Tennessee, USA

Correspondence to
Dr Jonathan L Haines, Center for Human Genetics Research, Vanderbilt University Medical Center, 519 Light Hall, 2215 Garland Avenue South, Nashville, TN 37232-0700, USA; jonathan@chgr.mc.vanderbilt.edu

Received 10 May 2013
Accepted 3 October 2013
Published Online First 22 October 2013

To cite: Davis MF, Sriram S, Bush WS, *et al*. *J Am Med Inform Assoc* 2013;**20**: e334–e340.

ABSTRACT

Objectives The clinical course of multiple sclerosis (MS) is highly variable, and research data collection is costly and time consuming. We evaluated natural language processing techniques applied to electronic medical records (EMR) to identify MS patients and the key clinical traits of their disease course.

Materials and methods We used four algorithms based on ICD-9 codes, text keywords, and medications to identify individuals with MS from a de-identified, research version of the EMR at Vanderbilt University. Using a training dataset of the records of 899 individuals, algorithms were constructed to identify and extract detailed information regarding the clinical course of MS from the text of the medical records, including clinical subtype, presence of oligoclonal bands, year of diagnosis, year and origin of first symptom, Expanded Disability Status Scale (EDSS) scores, timed 25-foot walk scores, and MS medications. Algorithms were evaluated on a test set validated by two independent reviewers.

Results We identified 5789 individuals with MS. For all clinical traits extracted, precision was at least 87% and specificity was greater than 80%. Recall values for clinical subtype, EDSS scores, and timed 25-foot walk scores were greater than 80%.

Discussion and conclusion This collection of clinical data represents one of the largest databases of detailed, clinical traits available for research on MS. This work demonstrates that detailed clinical information is recorded in the EMR and can be extracted for research purposes with high reliability.

INTRODUCTION

Patients with multiple sclerosis (MS) have a highly variable and poorly understood disease course, which varies from relatively minor intermittent and resolving neurological deficits to rapid, progressing, and permanent neurological deficits. Most research studies have focused on the origin of the disease, partly because of the difficulty in ascertaining sufficient longitudinal clinical data to study the disease course. Electronic medical records (EMR) may provide such a tool. We have previously shown that genomic signals of MS risk can be replicated using EMR-derived cohorts.[1 2] In this paper, we evaluated algorithms to identify patients with MS from the EMR and created new algorithms to extract detailed clinical information for the disease course of MS.

BACKGROUND AND SIGNIFICANCE

MS is a common, complex autoimmune disease with profound impact on the lives of individuals it touches. Despite rigorous study, much remains unknown about its pathophysiology and origins. Many genetic and environmental factors have been linked to the development of MS in an individual. In the past decade, scores of genetic variants have been associated to MS and replicated in subsequent studies.[3–5] Smoking, increased distance from the equator, and exposure to Epstein–Barr virus have been identified as risk factors.[6 7] While we do not fully understand how or why the disease develops, we know even less about the actual disease course, which is highly variable. Clinical data in EMR may be a rich resource of information that would allow greater research into the disease course of MS.

Clinical expression of the disease, including age of onset, rate of progression, and type and frequency of symptoms, varies drastically between individuals.[8–10] While genetic susceptibility to MS has been widely studied, there has been much less focus on its varied clinical expression. This is largely due to the difficulties and expense of collecting detailed longitudinal data on the large number of individuals often required for studies of complex diseases. However, these data are frequently recorded in physician notes typed or dictated into the EMR. While data recorded in medical records is less standardized than data collected expressly for research purposes, it is a rich resource that could be leveraged for complex diseases, such as MS.

Extracting data manually from medical records is tedious, time-consuming work that is prone to human error. The advent of EMR provides an opportunity to drastically shorten the time required to extract relevant medical information and decrease human error. Despite this promise, extracting information from EMR can be challenging. Typically, multimodal algorithms must be created, incorporating EMR components such as billing codes, medication data, laboratory values, and natural language processing (NLP) to achieve high positive predictive values (PPV) to identify disease states.[1 11 12] Identification of more detailed phenotypes, such as envisioned in 'next-generation phenotyping'[13] and drug response phenotypes, is more challenging and has only recently been explored.[14 15] We conducted a study of MS in a de-identified, research version of the EMR at Vanderbilt University Medical Center (VUMC) to determine the depth and range of clinical information relating to the disease course of MS in 5789 individuals.

MATERIALS AND METHODS

Study population

All medical records were obtained from the VUMC Synthetic Derivative—a research resource of over

two million de-identified records, including inpatients and outpatients.[1 16] Identifying information is removed from each record, including names, places, and identifying numbers, and the dates in each person's record are shifted consistently within a 365-day window in the past. The EMR at VUMC saw broad use as early as the 1990s, although not all clinical specialties adopted its use simultaneously. Relevant to this study, the MS Center at VUMC was established in 1994 and serves as both a primary and tertiary center for the evaluation and treatment of MS. The MS Center transitioned to computer-based documentation in 1997.

We utilized four previously published algorithms to identify MS patients from this database;[1] the algorithms focus on International Classification of Diseases, revision 9 (ICD-9) billing codes, prescribed MS treatments, and keywords located in the text. We made minor modifications, including increasing the number of ICD-9 codes for MS required in the 'definitive type 1' algorithm to require two or more instances and including the ICD-9 code for acute transverse myelitis (341.2) to the 'definitive type 2' and 'possible type 1' algorithms. These updated algorithms are publicly available on PheKB (http://www.phekb.org/phenotype/multiple-sclerosis-demonstration-project).

Algorithms to extract detailed clinical traits

Algorithms to extract clinical data from EMR text were implemented using Perl to access and search records stored in a MySQL database. Algorithms were initially developed using 899 records as a training dataset and then evaluated using a test set of 4890 records. Before algorithm development, we examined 60 training set records to determine what types of detailed clinical information related to the MS disease type and its course were often available, and how they were expressed in the clinical notes. We identified eight attributes: clinical subtype; presence of oligoclonal bands; year of diagnosis; Expanded Disability Status Scale (EDSS) score; timed 25-foot walk; year and origin of first neurological symptom; and MS medications. Our goal was to extract data explicitly stated in the medical record; we did not infer information (eg, the clinical subtype) from descriptions in the text.

Clinical subtype

The four clinical subtypes of MS are: relapsing remitting, secondary progressive, primary progressive and relapsing progressive. Subtypes were extracted from clinic notes, letters and problem lists (PL) that mentioned MS. Subtypes preceded or followed by words suggesting the clinician was not certain, such as 'questionable' or 'possible', were excluded by the use of regular expressions. As an individual may be classified with different subtypes over the course of their illness, all distinct subtypes mentioned for each individual were kept.

Oligoclonal bands

Over 85% of patients with MS have antibodies present in the cerebrospinal fluid and not in serum. These are referred to as oligoclonal bands and identifying these bands can aid clinicians in the diagnosis of MS.[17] As such testing is often performed by referring providers (and not repeated at referral centers, such as VUMC), it is important to search the clinical documentation in addition to laboratory results. We identified clinic notes, letters, and PL mentioning oligoclonal bands and extracted 200 characters surrounding the word 'oligoclonal'. The result was recorded as positive (ie, the clinician stated the test was positive or two or more bands were present) or negative (ie, the clinician stated the result was negative or no bands were observed) using regular expressions. No result was reported if one band was observed (inconclusive result). In the event that a person had both a negative and a positive result reported, the algorithm ignored the data and no conclusive result was recorded.

Year of diagnosis

MS is a clinically defined disease and the diagnostic criteria have evolved over the past 30 years.[18–20] Hence, the diagnosis of MS made by the clinician on a particular patient was based on the set of criteria that were relevant and operative at the time of the diagnosis. We extracted the year of diagnosis as recorded by the clinician, regardless of the definition used. Clinic notes and letters in the EMR were examined to identify mentions of the words 'diagnosis' and 'MS'. We identified exact, for example, '1975', and relative, for example, '3 years ago', dates that occurred within 70 characters of 'diagnosis'.

To determine the most likely diagnosis year, we first looked at exact references and recorded the most frequent year as the diagnosis year in our database. If no year of diagnosis was recorded in an exact reference, we analyzed relative references in the same manner. Identifying the most frequently reported year removed many typographical errors that were initially observed.

Measures of progression of disease disability

The EDSS[21] and timed 25-foot walk[22] are two measures used to monitor the progression of MS disability. Both can be recorded in structured fields in a manner similar to laboratory values. At VUMC, EDSS does not have a structured field but is often mentioned in clinic notes. The MS Center created a structured field for the timed 25-foot walk in 2008; however, scores have been collected and recorded in the text since 1999. We created algorithms to extract both of these measures from the narrative text in the absence of structured fields. Additional discussion of these measures and comparison of timed walk scores extracted from the clinical text and structured fields are included in the supplementary data (available online only).

The EDSS has a range from 0 (no disability due to MS) to 10 (death due to MS), in increments of 0.5.[21] The algorithm to extract these values from the text searched for 'EDSS' in notes, PL, and communications. Values (0–10) reported within 50 characters after 'EDSS' were extracted, and the closest number within range was recorded as EDSS scores.

To capitalize on the longitudinal aspect of timed 25-foot walks before structured values were available in 2008, we selected notes, then lines of text, from the clinical notes that mentioned 'timed walk', '25 feet', or '25 foot'. Times were extracted and recorded in seconds. The final output of this algorithm also noted if a walking aid (eg, cane) was mentioned.

Year and origin of first neurological symptom

As the clinical diagnosis of MS requires the presence of two lesions disseminated in space and time, patients are rarely diagnosed at the first presentation of neurological symptoms. However, the initial presentation of neurological symptoms of the disease may be important for research purposes and appears to aggregate in families (both the age and type of first neurological symptom).[23] While there are many references to symptoms in the narrative text, a complete neurological history must be investigated to be confident of identifying the first neurological symptom. We noticed that such a history was often reported in letters written from physicians at the MS Center to referring physicians and we restricted our algorithms to search these letters. The algorithm to identify the year of initial

neurological symptom selected 100 characters around phrases referencing the beginning of the disease course, that is, 'dating back' and 'began'. Specific dates were extracted from these phrases, either exact or relative.

To identify the type of first neurological symptom, 250 characters surrounding phrases that referenced the beginning of the disease course were extracted and run through the KnowledgeMap concept identifier,[24 25] which is a general purpose NLP system supporting negation and word-sense disambiguation, similar to MetaMap.[26] Concept unique identifiers (CUI) representing neurological symptoms were selected as the output of interest, as identified using Unified Medical Language System semantic types (see supplementary data, available online only). We then used text keywords and CUI to group the symptoms into central nervous system site of origin (brain stem, optic nerve, or spinal cord) using a list of MS-related neurological symptoms we compiled. Symptoms that did not fall into one of these categories were marked as 'other'. If more than one origin was identified, all were recorded and the origin was marked 'polysymptomatic'. Figure 1 provides a schematic of this algorithm.

Medications

Medications administered for the treatment of MS are fairly specific to this disease. MS medications are often discussed in a clinic visit with the patient and the patient is sent home with pamphlets to determine which medication they wish to start. Although VUMC has electronic prescribing tools, many outpatient prescriptions (especially in the early 2000s) are only documented in the free text of clinical notes, clinical messaging systems, or PL, and this has been especially true of the MS Center. Discussion of MS medications in narrative text could be because the patient is on the medication, the patient failed the medication due to continued progression of MS or excessive side effects, the clinician is considering the medication for the patient in the future, or the patient came into the clinic with questions regarding a specific treatment. To retrieve medications the patients were actually taking, we focused our efforts on extracting MS-related medications from PL only. The goal of this algorithm was to determine if a patient was ever on a medication. Extracted medications include interferon β-1a, interferon β-1b, glatiramer acetate, fingolimod, natalizumab, mitoxantrone, and teriflunomide. Text matching, using brand and generic names, was done in PL text to create a list of medications the patient had taken. Electronic prescribing tools automatically update the PL, so this method should also capture electronic prescriptions with near-perfect fidelity.

Evaluation

We reviewed the Synthetic Derivative records for 367 individuals across all case algorithms to create a gold standard for MS case status. Individuals were selected randomly within each MS algorithm type, and at least 50 individuals per case selection algorithm were reviewed. Each individual was categorized as diagnosed with MS, possible MS, or no MS, based on clinician impressions.

One hundred records were selected randomly from the test set for a blinded evaluation of the clinical trait algorithms. These records were reviewed manually for all clinical characteristics extracted by algorithms to define a gold standard. The reviewer recorded the information the treating clinician(s) appeared the most confident in by the end of the record (clinical subtype, year of diagnosis). The first 20 records were reviewed independently by author SS, a board-certified neurologist and founder and chief of the MS Center and a graduate student (MFD), with any discrepancies adjudicated by a second board-certified internist (JCD), blinded to the source of discrepancy. Given high initial concordance (92–100% per trait, median 99%) the graduate student performed the manual abstraction of the following 80 records. Kappa values were between 0.8 and

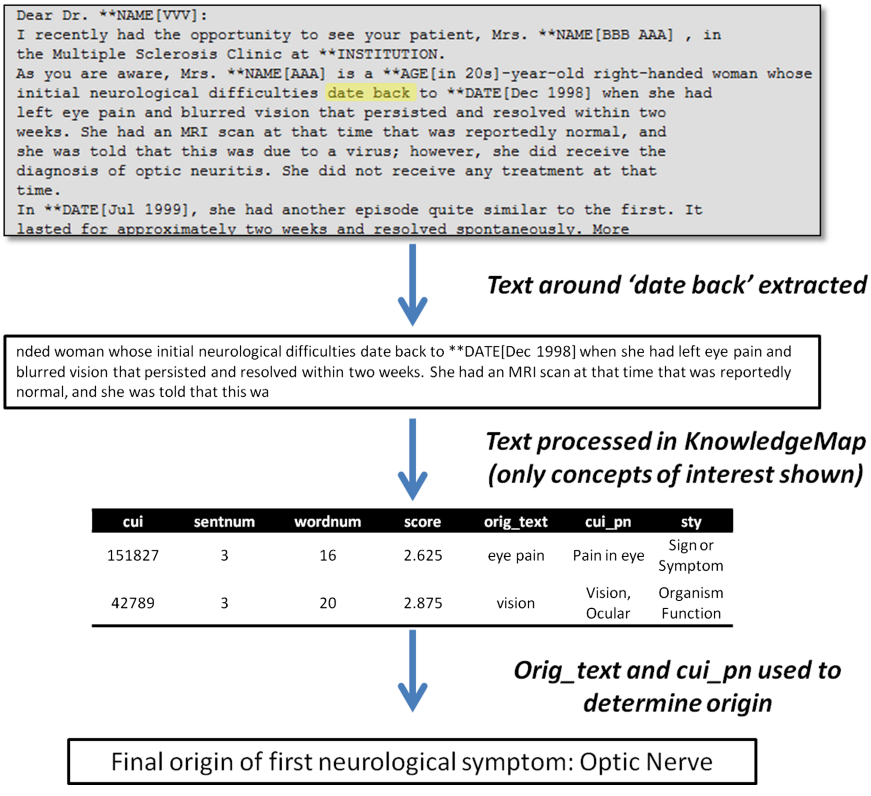

Figure 1 Schematic to represent how the algorithm to determine the origin of first neurological symptom works.

1.0 for each trait, with a median of 0.96. Manual abstraction took an average of 12.6 min per individual record, with a range of 1–40 min.

Demographic and clinical trait extraction data from the subset of records reviewed manually were compared to the overall dataset and demonstrated that the subset is an accurate representation of the overall dataset (see supplementary data, available online only).

PPV for case algorithms were calculated twice, with and without possible cases included as true positives. Clinical trait data derived from manual abstraction were compared to data extracted via the algorithms designed in this study. For all traits, recall, precision, specificity, and F-measure were calculated. True positives were defined as algorithm-extracted clinical traits that matched those found by manual abstraction; more than one true positive per person was possible for the EDSS, clinical subtype, timed 25-foot walk, medications, and origin of first symptom algorithms. We defined a person as a true negative when no values were extracted by the algorithm and when no values were also found by manual abstraction. There was thus a maximum of one true negative per trait per person.

Table 2 Demographics of all extracted cases

	No of individuals
Gender	
Female	4484
Male	1305
Age	
Median	54
Range	8–107
Deceased	508
Ethnicity	
White	3513
Black	440
Asian	11
Hispanic	16
Native American	1
Unknown	1808

Age is calculated for the year 2013 using birth year. Deceased includes individuals reported deceased in the EMR by linkage to the social security death index.
EMR, electronic medical record.

RESULTS

Case selection accuracy

A total of 5789 individuals was identified as cases by algorithms, with 4060 (70%) individuals matching one of the 'definite' criteria (table 1). PPV ranged from 16% to 96%. Reported demographics for all individuals are listed in table 2. Median follow-up time by individual was 4.5 years (range 0–20 years). On review, there were more false positives in the 'possible type 1' category than desired, including many individuals who were seen in the MS Center for other diseases, such as neurosarcoidosis.

Clinical trait extraction

Our algorithms extracted information for each clinical trait of interest in 903 (16%) to 3523 (61%) out of 5789 total MS individuals (table 3). Specificities for all algorithms were high, with seven of eight algorithms achieving specificity greater than 90% (table 4). Precision ranged from 87% to 99%. For clinical subtype and timed 25-foot walk, recall was at least 90%. However, recalls for year of diagnosis and origin of first symptom were 33% and 23%, respectively. The F-measure for all traits except year of diagnosis and origin of first symptom was above 70%.

After comparison to the gold standard was complete, we identified the need for minor changes in the algorithms for timed 25-foot walk, year of first symptom, and origin of first symptom, which significantly increased recall compared to the original algorithms at a nominal p value of 0.05 (p=0.02, 0.03, 0.02, respectively; table 5). During compilation into the database, some spaces and new lines were removed. We allowed for such changes by making spaces optional in regular expressions for timed walks and year and origin of first symptom. In addition, we identified another note title that represented letters to referring physicians and included the year and origin of first symptom. The F-measure for the algorithm of origin of first symptom also increased significantly (p=0.02).

DISCUSSION

We identified a large number of individuals with MS and detailed clinical information with minimal cost and time requirements. Both the MS case algorithms and the algorithms to extract detailed MS information performed well, with a precision between 87% and 100%. We are unaware of any other published dataset of MS patients of this size that has such detailed clinical information. This dataset provides a rich resource for better understanding MS and also shows that extraction of detailed disease states and markers of prognosis in patients with chronic disease is possible and may yield a powerful tool in chronic disease research.

While many studies have identified individuals serving as cases and controls for disease status from EMR,[1 11 27 28] this is

Table 1 Counts of individuals selected by case algorithms

Algorithm	No of samples	PPV* (%)	PPV† (%)
Definitive type 1	3975	96	96
Definitive type 2	85	64	79
Possible type 1	1315	16	64
Possible type 2	414	72	86
Total	5789	–	–

Algorithm details are available at http://www.phekb.org/phenotype/multiple-sclerosis-demonstration-project.
*Possible cases counted as false positives.
†Possible cases counted as true positives.
PPV, positive predictive value.

Table 3 Number of individuals for whom information was extracted for each clinical trait out of 5789

Clinical trait	Individuals, n
Clinical subtype	3140
Oligoclonal bands	1043
Year of diagnosis	1053
EDSS	903
Timed 25-foot walk	3523
Year of first symptom	2301
Origin of first symptom	1288
MS medications	2586

EDSS, Expanded Disability Status Scale; MS, multiple sclerosis.

Table 4 Statistics of algorithms compared to blinded manual review of 100 charts for all characteristics

Clinical trait	Gold standard positives, n*	Correctly identified, n*	Recall, %	Precision, %	Specificity, %	F-measure, %
Clinical MS subtype	61	60	98	88	81	93
Oligoclonal bands	28	20	71	87	97	78
Year of diagnosis	51	17	33	89	100	49
Expanded disability status scale	75	61	81	94	100	87
Timed 25-foot walk	120	99	83	99	100	90
Year of first symptom	56	24	43	100	100	60
Origin of first symptom	62	14	23	88	100	36
MS medications	99	63	64	95	93	76

*n refers to how many instances were recorded, not number of individuals. For EDSS, clinical subtype, timed 25-foot walk, medications, and origin of first symptom, this could be more than one per individual.
EDSS, Expanded Disability Status Scale; MS, multiple sclerosis.

one of the first studies to focus on specific clinical traits of a disease by text mining of the EMR. A few other studies have used text mining approaches to extract blood pressures, pacemaker implantations, and left ventricular ejection fractions as a marker of heart failure.[29–31] We have shown that detailed clinical information valuable to research studies is recorded in medical records of individuals with MS, and that this information can be extracted in a highly reliable manner. Such methods could potentially be applied across multiple EMR, such as envisioned by the eMERGE network [32] and SHRINE.[33]

We aimed for high precision to create a reliable database of information, rather than focusing on high recall, although the resulting recall of many algorithms was high. The ability to create highly specific algorithms for these clinical traits is due to many factors, many attributable to the nature of the disease studied. A diagnosis of MS is rarely given if a patient does not meet the criteria that are relatively specific to this disease, and diagnosis is generally verified by a neurologist. Treatments for MS are rarely used in other diseases. VUMC has a MS Center, with only five clinicians since its opening in 1997. This has resulted in a large number of clinic notes focused on the disease course of MS for each individual and much less variability in the style and content of clinic notes than may be found in other disease clinics. It should be noted, however, that not all individuals whose records we analyzed were enrolled in the MS Center or were even seen by a VUMC neurologist. These patients were likely to be seen at VUMC for other reasons and treated for MS elsewhere. While the 'possible type 1' algorithm identified a number of individuals with MS, the majority of individuals had not been definitively diagnosed. Depending on the purpose of the study, individuals identified by this algorithm should be used with caution.

Laboratory values are easily extractable via EMR, as each result is stored under the type of test done. However, the drawback to using EMR-derived data for laboratory values is that if the test was not performed at the primary institution (eg, VUMC), it will not be reported in a structured field. For example, the test for oligoclonal bands is most commonly ordered when trying to make a diagnosis of MS. Indeed, only 24% of cases had a value for oligoclonal bands in the relevant structured fields. Because this is a common test performed when diagnosing MS, the result is often echoed in the narrative text. We capitalized on clinic note references to extract this information in an additional group of individuals.

Structured fields in the EMR would also be the most accurate way to store and extract non-laboratory data, such as the EDSS and timed 25-foot walk measures. Unfortunately, these fields do not always contain the desired information due to the nature of the data or the EMR, and NLP provides an opportunity to recapture these data. We used NLP to extract timed 25-foot walk scores that were recorded before the existence of the structured field. Timed 25-foot walk scores derived from structured fields and NLP methods show no significant difference in our dataset (figure 2; see supplementary data, available online only), further validating NLP methods as a secondary means of data extraction.

Initially, we used MedEx[34] to extract medications but found it challenging to produce a medication list with high recall and PPV. To increase the likelihood that a medication mentioned represents one currently being taken, we required the presence of dosage and route information (extracted by MedEx). However, the majority of MS medications are given in one dose and one type of administration, so this information was often missing in the clinic record. Therefore, it was difficult to differentiate, without further NLP, if a medication was one being taken or being discussed for another reason. Because of these difficulties, we focused on the extraction of medications from PL, which contain active lists of medications for each patient. By doing this, we gained greater confidence in determining which medications a person had been taking. However, PL are not always updated, resulting in a lower recall rate than desired.

The algorithms we have written are not overly intricate, yet have yielded an extensive amount of clinical data on a large population. Additional work on these scripts could yield even

Table 5 Statistics of algorithms after additional modifications

Clinical trait	Gold standard positives, n*	Correctly identified, n*	Recall, %	Precision, %	Specificity, %	F-measure, %
Timed 25-foot walk	120	108	90	99	100	94
Year of first symptom	56	31	55	97	100	70
Origin of first symptom	62	21	34	88	93	49

*n refers to how many instances were recorded, not number of individuals. For timed 25-foot walk and origin of first symptom, this could be more than one per individual.

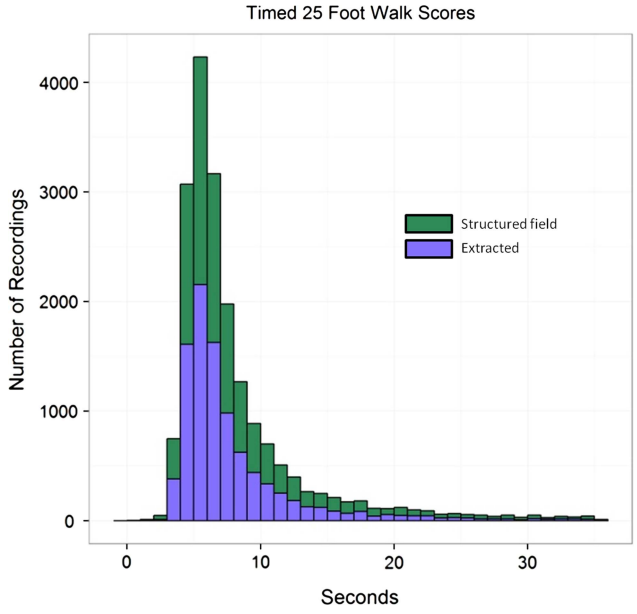

Figure 2 Distributions of timed 25-foot walk scores as found in the structured fields and extracted from the text of the clinical records.

greater recall for the clinical traits studied here, and it is likely that other clinical traits could be extracted. For example, we did not attempt to extract information about the number, length, or types of relapses experienced by individuals or start and stop dates of medications. In future research, we also hope to extract reasons for why medications are halted—ineffectiveness, unsustainable side effects, patient non-compliance, etc. The scripts described in this paper searched for specific references by the clinician about clinical traits. They did not use the text to infer information, such as diagnosis year or clinical subtype, both of which could have been done to enhance recall. In particular, we had very low recall in our algorithm to extract diagnosis year. On review of instances of algorithm failure, many times we missed when a patient was diagnosed in the course of the record, as it is rare that a clinician would record the current year, instead stating, 'I believe Mr. [NAME] fully meets the criteria for a diagnosis of MS' or simply listing 'MS' as the final impression of the clinic visit. Algorithms targeting current diagnoses would greatly improve the recall of this clinical trait.

The application of the subject selection and clinical trait algorithms proved to be great tools in the creation of a large dataset of MS individuals with longitudinal disease course data at VUMC. Further use of these algorithms would be to apply them to EMR datasets in other institutions. The subject selection algorithms should be easily transferable as there are no parts of the algorithm that are specific to VUMC records. The transferability of the clinical trait algorithms is likely to vary. We expect the most difficult algorithms to transfer would be the age and type of first neurological symptom, which rely on clinician-specific wording to identify referral letters that contain a history with specific key words. The general principle could be carried over but evaluation of the clinic notes should be done to evaluate the format of the notes at the intended university or clinic. The presence of oligoclonal bands and timed 25-foot walk algorithms rely on no institution-specific formats. Ascertainment of structured fields at any institution should first be attempted; however, the ease with which we were able to identify these scores suggests NLP-derived algorithms would work well at other institutions if needed. Additional methods of detecting the results in the text could be added if deemed necessary. For instance, abbreviations for the timed walk, including 'ft' and 'T25FW', were not seen in the records we reviewed but they may be used at other institutions. We know of no specific reasons why the algorithms for age at diagnosis, EDSS, and clinical subtype would not be transferable. The algorithm for medications would depend on the existence of PL at the institution of interest.

CONCLUSIONS

EMR databases are a rich resource of detailed information of the clinical course of MS. This information is extractable from clinic notes by simple algorithms, with high specificity, precision, and recall.

Acknowledgements The authors would like to thank the patients at VUMC for providing them with this research opportunity. They would like to thank the Vanderbilt multiple sclerosis clinic for shedding further light on the EMR at VUMC and common clinic practices.

Contributors JLH, MFD, and SS conceived the study design. MFD, JCD, and WSB were involved in creation of the algorithms. All authors contributed to the manuscript writing and approved the final manuscript.

Funding This work was supported by grants NS032830 (to JLH), LM010685 (to JCD), P32GM080178 (to Vanderbilt University) and UL1TR000445 (to Vanderbilt CTSA). The contents of this paper are solely the responsibility of the authors and do not necessarily represent the official views of the National Center for Advancing Translation Sciences or the National Institutes of Health.

Competing interests None.

Provenance and peer review Not commissioned; externally peer reviewed.

REFERENCES

1 Ritchie MD, Denny JC, Crawford DC, *et al*. Robust replication of genotype-phenotype associations across multiple diseases in an electronic medical record. *Am J Hum Genet* 2010;86:560–72.
2 Denny JC, Ritchie MD, Basford MA, *et al*. PheWAS: demonstrating the feasibility of a phenome-wide scan to discover gene-disease associations. *Bioinformatics* 2010;26:1205–10.
3 Naito S, Namerow N, Mickey MR, *et al*. Multiple sclerosis: association with HL-A3. *Tissue Antigens* 1972;2:1–4.
4 Gregory SG, Schmidt S, Seth P, *et al*. Interleukin 7 receptor alpha chain (IL7R) shows allelic and functional association with multiple sclerosis. *Nat Genet* 2007;39:1083–91.
5 Sawcer S, Hellenthal G, Pirinen M, *et al*. Genetic risk and a primary role for cell-mediated immune mechanisms in multiple sclerosis. *Nature* 2011;476:214–19.
6 Ascherio A, Munger KL. Environmental risk factors for multiple sclerosis. Part I: the role of infection. *Ann Neurol* 2007;61:288–99.
7 Ebers GC. Environmental factors and multiple sclerosis. *Lancet Neurol* 2008;7:268–77.
8 Sadovnick AD. European Charcot Foundation Lecture: the natural history of multiple sclerosis and gender. *J Neurol Sci* 2009;286:1–5.
9 Confavreux C, Vukusic S, Moreau T, *et al*. Relapses and progression of disability in multiple sclerosis. *N Engl J Med* 2000;343:1430–8.
10 Runmarker B, Andersen O. Prognostic factors in a multiple sclerosis incidence cohort with twenty-five years of follow-up. *Brain* 1993;116:117–34.
11 Kho AN, Pacheco JA, Peissig PL, *et al*. Electronic medical records for genetic research: results of the eMERGE consortium. *Sci Transl Med* 2011;3:79re1–1.
12 Denny JC. Chapter 13: mining electronic health records in the genomics era. *PLoS Comput Biol* 2012;8:e1002823.
13 Hripcsak G, Albers DJ. Next-generation phenotyping of electronic health records. *J Am Med Inform Assoc* 2013;20:117–21.
14 Lependu P, Iyer SV, Bauer-Mehren A, *et al*. Pharmacovigilance using clinical notes. *Clin Pharmacol Ther* 2013;93:547–55.
15 Delaney JT, Ramirez AH, Bowton E, *et al*. Predicting clopidogrel response using DNA samples linked to an electronic health record. *Clin Pharmacol Ther* 2012;91:257–63.

16 Roden DM, Pulley JM, Basford MA, *et al*. Development of a large-scale de-identified DNA biobank to enable personalized medicine. *Clin Pharmacol Ther* 2008;84:362–9.
17 Dobson R, Ramagopalan S, Davis A, *et al*. Cerebrospinal fluid oligoclonal bands in multiple sclerosis and clinically isolated syndromes: a meta-analysis of prevalence, prognosis and effect of latitude. *J Neurol Neurosurg Psychiatry* 2013; 84:909–14.
18 Poser CM, Paty DW, Scheinberg L, *et al*. New diagnostic criteria for multiple sclerosis: guidelines for research protocols. *Ann Neurol* 1983;13:227–31.
19 McDonald WI, Compston A, Edan G, *et al*. Recommended diagnostic criteria for multiple sclerosis: guidelines from the International Panel on the diagnosis of multiple sclerosis. *Ann Neurol* 2001;50:121–7.
20 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.
21 Kurtzke JF. Rating neurologic impairment in multiple sclerosis: an expanded disability status scale (EDSS). *Neurology* 1983;33:1444–52.
22 Fischer JS, Rudick RA, Cutter GR, *et al*. The Multiple Sclerosis Functional Composite Measure (MSFC): an integrated approach to MS clinical outcome assessment. National MS Society Clinical Outcomes Assessment Task Force. *Mult Scler* 1999;5:244–50.
23 Barcellos LF, Oksenberg JR, Green AJ, *et al*. Genetic basis for clinical expression in multiple sclerosis. *Brain* 2002;125:150–8.
24 Denny JC, Spickard A III, Miller RA, *et al*. Identifying UMLS concepts from ECG Impressions using KnowledgeMap. *AMIA Annu Symp Proc* 2005;2005:196–200.
25 Denny JC, Smithers JD, Miller RA, *et al*. "Understanding" medical school curriculum content using KnowledgeMap. *J Am Med Inform Assoc* 2003;10:351–62.
26 Aronson AR, Lang FM. An overview of MetaMap: historical perspective and recent advances. *J Am Med Inform Assoc* 2010;17:229–36.
27 Perlis RH, Iosifescu DV, Castro VM, *et al*. Using electronic medical records to enable large-scale studies in psychiatry: treatment resistant depression as a model. *Psychol Med* 2012;42:41–50.
28 Ananthakrishnan AN, Cai T, Savova G, *et al*. Improving case definition of Crohn's disease and ulcerative colitis in electronic medical records using natural language processing: a novel informatics approach. *Inflamm Bowel Dis* 2013;19:1411–20.
29 Turchin A, Kolatkar NS, Grant RW, *et al*. Using regular expressions to abstract blood pressure and treatment intensification information from the text of physician notes. *J Am Med Inform Assoc* 2006;13:691–5.
30 Rosier A, Burgun A, Mabo P. Using regular expressions to extract information on pacemaker implantation procedures from clinical reports. *AMIA Annu Symp Proc* 2008;2008:81–5.
31 Garvin JH, DuVall SL, South BR, *et al*. Automated extraction of ejection fraction for quality measurement using regular expressions in Unstructured Information Management Architecture (UIMA) for heart failure. *J Am Med Inform Assoc* 2012;19:859–66.
32 Gottesman O, Kuivaniemi H, Tromp G, *et al*. The Electronic Medical Records and Genomics (eMERGE) Network: past, present, and future. *Genet Med* 2013;15:761–71.
33 Weber GM, Murphy SN, McMurry AJ, *et al*. The Shared Health Research Information Network (SHRINE): a prototype federated query tool for clinical data repositories. *J Am Med Inform Assoc* 2009;16:624–30.
34 Xu H, Stenner SP, Doan S, *et al*. MedEx: a medication information extraction system for clinical narratives. *J Am Med Inform Assoc* 2010;17:19–24.