IDENTIFYING NEUROMYELITIS OPTICA PATIENTS FROM INSURANCE CLAIMS DATA USING NFERX, A NATURAL LANGUAGE PROCESSING-BASED PLATFORM

Authors: Enrique Garcia-Rivera[1], Jiho Park[1], Zainab Doctor[1], Agustin Lopez-Marquez[1], Danny Sheinson[2], Craig S. Meyer[2], Tu-My To[2]
Affiliations: [1]nference, [2]Genentech

BACKGROUND

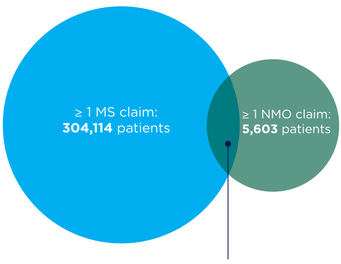
Neuromyelitis optica spectrum disorder (NMO) is a rare auto-immune disease affecting the central nervous system and has clinical characteristics similar to multiple sclerosis (MS), making clinical diagnosis challenging. This study aims to improve the identification of NMO patients by leveraging nferX, a natural language processing platform, to enable encoding of patient claims that are specific to NMO versus MS.

METHODS

Patient Selection

Using the IQVIA Pharmetrics Plus database, we defined two separate patient cohorts by MS or NMO: one cohort of 304,114 patients with ≥ 1 MS claim (ICD-9 diagnosis code 340 or ICD-10 diagnosis code G35) and another cohort of 5,603 patients with ≥ 1 NMO claim (ICD-9 diagnosis code 341.0 or ICD-10 diagnosis code G36.0). 2,551 patients were overlapping between the two cohorts, with ≥ 1 MS and NMO claim **(Figure 1).**

Figure 1



Patient Encoding

For all selected patients, all non-MS and non-NMO claims were encoded into a patient matrix. To encode the claims data for each patient, we first removed the ICD diagnosis codes that corresponded to MS or NMO. We then narrowed the set of codes considered to a subset of differential codes – defined as the ICD diagnosis, ICD procedure, and NDC drug codes that were present in ≥ 0.1% of patients in both cohorts and had a ≥ 2-fold prevalence enrichment in either cohort. In the resulting patient-to-code matrix **(Figure 2)**, each row corresponds to a unique patient, and each column corresponds to a unique code. Each entry corresponds to the number of unique dates that a given code appeared on a claim for that patient. We examined scenarios considering only ICD codes, only NDC codes, both, and at varying time windows relative to the first identified MS or NMO code (within 1 year post diagnosis and all years of data).

Figure 2

Restrict codes to a single occurrence per day

Patient	Claim Date	Code	Differential Code?
A.B.C.	Feb-1-2015	ICD-12	Yes
A.B.C.	Feb-1-2015	ICD-12	Yes
A.B.C.	Feb-1-2015	NDC-12	Yes
A.B.C.	Apr-1-2015	ICD-12	Yes
A.B.C.	Apr-1-2015	ICD-34 (MS)	Yes
A.B.C.	Apr-1-2015	ICD-56	Yes
A.B.C.	Apr-1-2015	NDC-34	No
A.B.C.	Apr-1-2015	NDC-34	No
...	...	...	...
XYZ.	Mar-1-2014	ICD-78 (NMO)	Yes
XYZ.	Mar-1-2014	ICD-90	No
XYZ.	Aug-1-2014	ICD-90	No
XYZ.	Aug-1-2014	NDC-12	Yes
XYZ.	Nov-1-2014	NDC-56	Yes

Remove duplicate codes in the same day

Patient	Claim Date	Code	Differential Code?
A.B.C.	Feb-1-2015	ICD-12	Yes
A.B.C.	Feb-1-2015	NDC-12	Yes
A.B.C.	Apr-1-2015	ICD-12	Yes
A.B.C.	Apr-1-2015	ICD-34 (MS)	Yes
A.B.C.	Apr-1-2015	ICD-56	Yes
A.B.C.	Apr-1-2015	NDC-34	No
...	...	...	...
XYZ.	Mar-1-2014	ICD-78 (NMO)	Yes
XYZ.	Aug-1-2014	ICD-90	No
XYZ.	Aug-1-2014	NDC-12	Yes
XYZ.	Nov-1-2014	NDC-56	Yes

Remove ICD diagnosis codes that correspond to MS or NMO

Patient	Claim Date	Code	Differential Code?
A.B.C.	Feb-1-2015	ICD-12	Yes
A.B.C.	Feb-1-2015	NDC-12	Yes
A.B.C.	Apr-1-2015	ICD-12	Yes
A.B.C.	Apr-1-2015	ICD-34 (MS)	Yes
A.B.C.	Apr-1-2015	ICD-56	Yes
A.B.C.	Apr-1-2015	NDC-34	No
...	...	...	...
XYZ.	Mar-1-2014	ICD-78 (NMO)	Yes
XYZ.	Mar-1-2014	ICD-90	No
XYZ.	Aug-1-2014	ICD-90	No
XYZ.	Aug-1-2014	NDC-12	Yes
XYZ.	Nov-1-2014	NDC-56	Yes

Remove MS codes and NMO codes

Patient	Claim Date	Code	Differential Code?
A.B.C.	Feb-1-2015	ICD-12	Yes
A.B.C.	Feb-1-2015	NDC-12	Yes
A.B.C.	Apr-1-2015	ICD-12	Yes
A.B.C.	Apr-1-2015	ICD-56	Yes
A.B.C.	Apr-1-2015	NDC-34	No
...	...	...	...
XYZ.	Mar-1-2014	ICD-90	No
XYZ.	Aug-1-2014	ICD-90	No
XYZ.	Aug-1-2014	NDC-12	Yes
XYZ.	Nov-1-2014	NDC-56	Yes

Keep only the subset of differential codes

Patient	Claim Date	Code	Differential Code?
A.B.C.	Feb-1-2015	ICD-12	Yes
A.B.C.	Feb-1-2015	NDC-12	Yes
A.B.C.	Apr-1-2015	ICD-12	Yes
A.B.C.	Apr-1-2015	ICD-56	Yes
A.B.C.	Apr-1-2015	NDC-34	No
...	...	...	...
XYZ.	Mar-1-2014	ICD-90	No
XYZ.	Aug-1-2014	ICD-90	No
XYZ.	Aug-1-2014	NDC-12	Yes
XYZ.	Nov-1-2014	NDC-56	Yes

Only keep differential codes

Codes with ≥ 0.1% prevalence in both cohorts and ≥ 2-fold prevalence enrichment in either cohort

Patient	Claim Date	Code	Differential Code?
A.B.C.	Feb-1-2015	ICD-12	Yes
A.B.C.	Feb-1-2015	NDC-12	Yes
A.B.C.	Apr-1-2015	ICD-12	Yes
A.B.C.	Apr-1-2015	ICD-56	Yes
XYZ.	Aug-1-2014	NDC-12	Yes
XYZ.	Nov-1-2014	NDC-56	Yes

Determine code frequencies on a per-patient basis

Patient	Claim Date	Code	Differential Code?
A.B.C.	Feb-1-2015	ICD-12	Yes
A.B.C.	Feb-1-2015	NDC-12	Yes
A.B.C.	Apr-1-2015	ICD-12	Yes
A.B.C.	Apr-1-2015	ICD-56	Yes
...	...	...	...
XYZ.	Aug-1-2014	NDC-12	Yes
XYZ.	Nov-1-2014	NDC-56	Yes

Get code frequencies per patient

Patient	Code	Code Frequency
A.B.C.	ICD-12	2
A.B.C.	ICD-56	1
A.B.C.	NDC-12	1
...	...	...
XYZ.	NDC-12	1
XYZ.	NDC-56	1

Transforming to nferX Phenotypes

This patient-code matrix was then transformed to a patient-concept matrix **(Figure 3)** utilizing the nferX Platform. nferX, a software platform developed by nference, leverages natural language processing approaches to quantify significant text associations from 110+ million publicly available documents. For each ICD code, we mapped a set of associated disease phenotypes from the code description using nferX. For example, we extracted the following disease phenotypes from ICD-10 diagnosis code H4713 ("papilledema associated with retinal disorder"): "retinal diseases", "papilledema associated with retinal disorder", and "papilledema". For each patients, the code occurrences were summed and assigned to the relevant mapped concepts.

As a final step, we removed any concepts that contained any of the following MS-related or NMO-related terms: "multiple sclerosis", "RRMS", "PPMS", "relapsing remitting", "primary progressive", "neuromyelitis optica" and "NMO". In the resulting patient to concept matrix **(Figure 3),** each row corresponds to a unique patient, and each column corresponds to a unique concept (e.g. disease phenotype or drug). Each entry corresponds to the number of times a given concept occurred for a patient during the time period of interest.

Figure 3

Map each code to its associated concepts using nferX (phenotype or drug)

Patient	Code	Code Frequency
A.B.C.	ICD-12	2
A.B.C.	ICD-56	1
A.B.C.	NDC-12	1
...	...	...
XYZ.	NDC-12	1
XYZ.	NDC-56	1

Get associated concepts (phenotype or drug)

Patient	Code	Code Frequency	Associated Concept
A.B.C.	ICD-12	2	Phenotype-1
A.B.C.	ICD-12	2	Phenotype-2
A.B.C.	ICD-12	2	Phenotype-3
A.B.C.	ICD-12	2	Phenotype-4
A.B.C.	ICD-56	1	Phenotype-3
A.B.C.	ICD-56	1	Phenotype-5
A.B.C.	NDC-34	1	Drug-1
...	...	...	...
XYZ.	NDC-34	1	Drug-1
XYZ.	NDC-78	1	Drug-2

Encode the patient-phenotype matrix using the sum occurrences of the associated concepts

Patient	Code	Code Frequency	Associated Concept
A.B.C.	ICD-12	2	Phenotype-1
A.B.C.	ICD-12	2	Phenotype-2
A.B.C.	ICD-12	2	Phenotype-3
A.B.C.	ICD-12	2	Phenotype-4
A.B.C.	ICD-56	1	Phenotype-3
A.B.C.	ICD-56	1	Phenotype-5
A.B.C.	NDC-34	1	Drug-1
...	...	...	...
XYZ.	NDC-34	1	Drug-1
XYZ.	NDC-78	1	Drug-2

Sum occurrences of associated concepts

Patient	Associated Concept	Concept Frequency
A.B.C.	Phenotype-1	3
A.B.C.	Phenotype-2	2
A.B.C.	Phenotype-3	2
A.B.C.	Phenotype-4	2
A.B.C.	Phenotype-5	1
A.B.C.	Drug-1	1
...	...	...
XYZ.	Drug-1	1
XYZ.	Drug-2	1

Literature Weighting

We weighted each entry in the patient to concept matrix using the literature specificity of the entry's associated concept to NMO over MS using nferX's formulation of the pointwise mutual information (PMI) metric between concepts A and B, pmi_{AB} **(Equations 1-3)**. In this formulation, we defined the local context of a concept as the five concepts (e.g. individual words or phrases) immediately preceding and following every occurrence of that concept. The adjacency adj_{AB}, between two concepts A and B is the number of times concept A is found in concept B's local context, or vice-versa. We weighted each entry in the patient to concept matrix by the difference between the associated concept's PMI to NMO and PMI to MS. Finally, the row sum of the literature-weighted concept counts was calculated per patient. We defined a patient's summary score as the row sum, which captures the extent to which a patient's claims history is associated to NMO over MS in the literature **(Figure 4).**

Figure 4

Weight associated concepts by literature specificity to NMO over MS

Patient	Associated Concept	Concept Frequency
A.B.C.	Phenotype-1	3
A.B.C.	Phenotype-2	2
A.B.C.	Phenotype-3	2
A.B.C.	Phenotype-4	2
A.B.C.	Phenotype-5	1
A.B.C.	Drug-1	1
...	...	...
XYZ.	Drug-1	1
XYZ.	Drug-2	1

Weight associated concepts by literature specificity

Patient	Associated Concept	Concept Frequency	Literature Specificity	Weighted Value
A.B.C.	Phenotype-1	3	3.0	9.0
A.B.C.	Phenotype-2	2	-1.0	-2.0
A.B.C.	Phenotype-3	2	1.5	3.0
A.B.C.	Phenotype-4	2	1.0	2.0
A.B.C.	Phenotype-5	1	-4.0	-4.0
A.B.C.	Drug-1	1	-0.5	-0.5
...	...	...	...	...
XYZ.	Drug-1	1	-0.5	-0.5
XYZ.	Drug-2	1	2.5	2.5

Calculate the row sum for each patient

Patient	Associated Concept	Concept Frequency	Literature Specificity	Weighted Value
A.B.C.	Phenotype-1	3	3.0	9.0
A.B.C.	Phenotype-2	2	-1.0	-2.0
A.B.C.	Phenotype-3	2	1.5	3.0
A.B.C.	Phenotype-4	2	1.0	2.0
A.B.C.	Phenotype-5	1	-4.0	-4.0
A.B.C.	Drug-1	1	-0.5	-0.5
...	...	...	...	...
XYZ.	Drug-1	1	-0.5	-0.5
XYZ.	Drug-2	1	2.5	2.5

Compute row sum

Patient	Row Sum
A.B.C.	7.5
...	...
XYZ.	2.0

Equations 1-3

$$pmi_{AB} = log_{10}\left(\frac{P(A,B)}{P(A)\cdot P(B)}\right) \quad (1)$$

$$pmi_{AB} = log_{10}\left(\frac{\frac{adj_{AB}}{N_C}}{\frac{N_A}{N_C}\cdot\frac{N_B}{N_C}}\right) \quad (2)$$

$$pmi_{AB} = log_{10}\left(\frac{adj_{AB}\cdot N_C}{N_A\cdot N_B}\right) \quad (3)$$

Definition of nferX pointwise mutual information (PMI) metric.

The PMI between two concepts is defined as a function of: $P(A,B)$ – the probability of observing concepts A and B in each other's local context; $P(A)$ – the probability of observing concept A; $P(B)$ – the probability of observing concept B; N_A – the number of occurrences of concept A; N_B – the number of occurrences of concept B; and N_C – the summed number of occurrences of all concepts. All probabilities and occurrences are calculated for a corpus of interest.

Methods – Benchmarking the Approach

We analyzed the distribution of summary scores for various cohorts of high-confidence MS and NMO patients and quantified the effect size with the Cohen's D metric. We used two different approaches to define these cohorts of high-confidence MS and NMO patients **(Table 1).**

In our first approach, we used published algorithms to select MS and NMO patients (Culpepper *et al.,* 2019 and Ajmera *et al.*, 2018, respectively) as a guideline to define high-confidence patient cohorts. For the high-confidence MS cohort, we selected patients with more than 3 MS-related claims (ICD diagnosis for MS or a disease-modifying therapy) within one year. For the high-confidence NMO cohort, we selected patients with either one inpatient visit or two outpatient visits for an NMO diagnosis, or at least 1 NMO diagnosis together with a second diagnosis for either transverse myelitis or optic neuritis. We then excluded patients who were treated with beta-interferons or other MS-related treatments, or received an MS diagnosis after the initial NMO, transverse myelitis, or optic neuritis diagnosis.

In our second approach, we used the number of MS/NMO claims on a per patient basis to define high-confidence cohorts. For each disease, we selected all patients whose number of claims for that disease exceeded a given cutoff. We explored the following four options for this cutoff: 2, 10, 20, and 100 claims. Because an individual patient can have multiple claims for both diseases, it is important to note that the two high-confidence cohorts are not mutually exclusive.

Table 1

Counts from the high-confidence MS and NMO cohorts as defined by published algorithms or the number of MS/NMO claims on a per patient basis

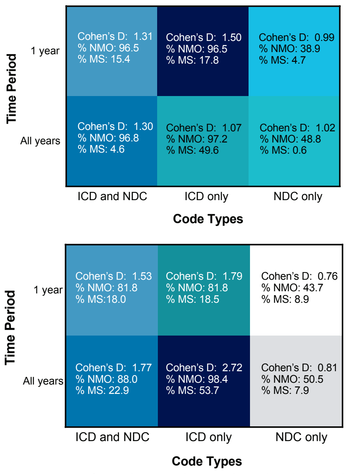
Methods	MS Cohort	NMO Cohort	Other Cohort
Published algorithms	**111,769**	**3,146**	**192,201**
≥ 2 disease claims	252,216	3,783	53,134
≥ 10 disease claims	138,686	1,471	167,560
≥ 20 disease claims	89,803	1,009	216,672
≥ 100 disease claims	15,802	226	291,181

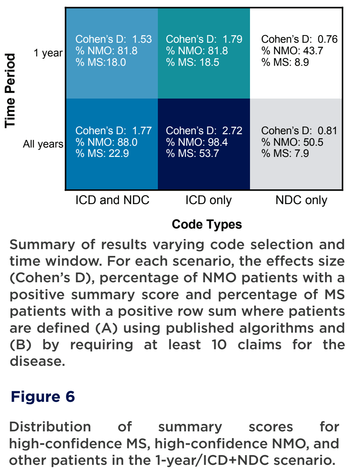
RESULTS

Positive values of pmi_{AB} correspond to a stronger literature association to NMO over MS. Thus, if the approach works, we'd expect patients with positive summary scores to come from the high-confidence NMO cohort, rather than the high-confidence MS cohort. We accordingly calculated the percentage of high-confidence MS patients who had positive summary scores, and the percentage of high-confidence NMO patients who had positive summary scores **(Figure 5)**. We found that the percentage of high-confidence NMO patients with positive summary scores was consistently high (between 96.5% and 97.2%) across the ICD+NDC and ICD-only scenarios. In contrast, the NDC only scenario showed a relatively low percentage of high-confidence NMO patients with positive summary scores (38.9% and 48.8%). In addition, the percentage of high-confidence MS patients with positive summary scores was minimized in the all years/NDC-only (0.60%) and all years/ICD+NDC (4.6%) scenarios.

In the scenario combining ICD+NDC codes and restricting analysis to 1 year after the initial diagnosis, Cohen's D separation in summary scores between literature-defined high-confidence NMO and MS cohorts was 1.31 **(Figure 5)**, with mean summary scores of 9.6 and -3.0 for the NMO and MS cohorts, respectively **(Figure 6).**

Figure 5

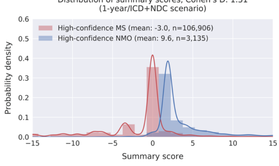




Summary of results varying code selection and time window. For each scenario, the effects size (Cohen's D), percentage of NMO patients with a positive summary score and percentage of MS patients with a positive row sum where patients are defined (A) using published algorithms and (B) by requiring at least 10 claims for the disease.

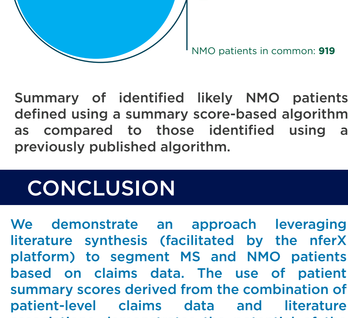
Figure 6

Distribution of summary scores for high-confidence MS, high-confidence NMO, and other patients in the 1-year/ICD+NDC scenario.



Finally, we utilized the summary scores derived from the literature-weighted patient concept matrix (1 year/ICD-only scenario) to identify a candidate NMO population. By applying a cutoff of one standard deviation above the mean summary score of MS patients, 14,822 candidate NMO patients were identified **(Figure 7)**. 13,150 patients without any claims for NMO were included that would not have otherwise been identified. Of the likely NMO patients identified using summary scores, 919 were in common with the 3,146 NMO patients selected using the published NMO algorithm.

Figure 7



Summary of identified likely NMO patients defined using a summary score-based algorithm as compared to those identified using a previously published algorithm.

CONCLUSION

We demonstrate an approach leveraging literature synthesis (facilitated by the nferX platform) to segment MS and NMO patients based on claims data. The use of patient summary scores derived from the combination of patient-level claims data and literature associations demonstrates the potential of the approach through separation of patient cohorts defined using previously-published algorithms. By applying a cutoff of one standard deviation above the mean summary score of MS patients, additional patients that otherwise would not have been identified as NMO patients were included. Further validation of this approach is needed to support its future use, both towards identifying NMO patients and its potential to identify patient populations for other diseases.