Using an artificial intelligence tool incorporating natural language processing to identify patients with a diagnosis of ANCA-associated vasculitis in electronic health records

Leeuwen, J.R. van; Penne, E.L.; Rabelink, T.; Knevel, R.; Teng, Y.K.O.

Computers in Biology and Medicine 168 (2024) 107757

Contents lists available at ScienceDirect

Computers in Biology and Medicine

journal homepage: www.elsevier.com/locate/compbiomed

Using an artificial intelligence tool incorporating natural language processing to identify patients with a diagnosis of ANCA-associated vasculitis in electronic health records

Jolijn R. van Leeuwen [a], Erik L. Penne [b], Ton Rabelink [a], Rachel Knevel [c], Y.K. Onno Teng [a,*]

[a] Center of Expertise for Lupus-, Vasculitis- and Complement-mediated Systemic diseases (LuVaCs), Department of Internal Medicine - Nephrology Section, Leiden University Medical Center, Leiden, the Netherlands
[b] Department of Internal Medicine – Nephrology Section, Northwest Clinics, Alkmaar, the Netherlands
[c] Department of Rheumatology, Leiden University Medical Center, Leiden, the Netherlands

ARTICLE INFO

Keywords:
ANCA-Associated vasculitis
Electronic-health-records
Artificial intelligence
Natural language processing
Pauci-immune glomerulonephritis

ABSTRACT

Background: Because anti-neutrophil cytoplasmatic antibody (ANCA)-associated vasculitis (AAV) is a rare, life-threatening, auto-immune disease, conducting research is difficult but essential. A long-lasting challenge is to identify rare AAV patients within the electronic-health-record (EHR)-system to facilitate real-world research. Artificial intelligence (AI)-search tools using natural language processing (NLP) for text-mining are increasingly postulated as a solution.

Methods: We employed an AI-tool that combined text-mining with NLP-based exclusion, to accurately identify rare AAV patients within large EHR-systems (>2.000.000 records). We developed an identification method in an academic center with an established AAV-training set (n = 203) and validated the method in a non-academic center with an AAV-validation set (n = 84). To assess accuracy anonymized patient records were manually reviewed.

Results: Based on an iterative process, a text-mining search was developed on disease description, laboratory measurements, medication and specialisms. In the training center, 608 patients were identified with a sensitivity of 97.0 % (95%CI [93.7, 98.9]) and positive predictive value (PPV) of 56.9 % (95%CI [52.9, 60.1]). NLP-based exclusion resulted in 444 patients increasing PPV to 77.9 % (95%CI [73.7, 81.7]) while sensitivity remained 96.3 % (95%CI [93.8, 98.0]). In the validation center, text-mining identified 333 patients (sensitivity 97.6 % (95%CI [91.6, 99.7]), PPV 58.2 % (95%CI [52.8, 63.6])) and NLP-based exclusion resulted in 223 patients, increasing PPV to 86.1 % (95%CI [80.9, 90.4]) with 98.0 % (95%CI [94.9, 99.4]) sensitivity. Our identification method outperformed ICD-10-coding predominantly in identifying MPO+ and organ-limited AAV patients.

Conclusions: Our study highlights the advantages of implementing AI, notably NLP, to accurately identify rare AAV patients within large EHR-systems and demonstrates the applicability and transportability. Therefore, this method can reduce efforts to identify AAV patients and accelerate real-world research, while avoiding bias by ICD-10-coding.

1. Introduction

Anti-neutrophil cytoplasmatic antibody (ANCA) associated vasculitis (AAV) is a rare, potentially life-threatening, systemic auto-immune disease [1,2]. AAV includes three sub-types of small vessel vasculitis: granulomatosis with polyangiitis (GPA, Wegener's), microscopic polyangiitis (MPA) and eosinophilic GPA (EGPA, Churg–Strauss syndrome). Give the high probability of relapses, accompanied organ-damage and heterogeneity between patients, research is essential to further improve and personalize treatments [1,2]. Assembling a relevant cohort size for clinical research is difficult for rare diseases, including AAV, making improvement and innovations slow. There is increasing awareness on the added value of real-world data, captured in electronic health records or databases, to improve insight into epidemiology, assess long-term outcomes efficacy and safety [3–5]. Advantages of real-world data are that large cohorts can be analyzed relatively fast and will include

* Corresponding author.Department of Nephrology, Leiden University Medical Center (LUMC), P.O. Box 9600, 2300 RC Leiden, the Netherlands.
E-mail address: Y.K.O.Teng@lumc.nl (Y.K.O. Teng).

https://doi.org/10.1016/j.compbiomed.2023.107757
Received 7 September 2023; Received in revised form 14 November 2023; Accepted 21 November 2023
Available online 25 November 2023

patients often excluded from clinical trials due to age, pregnancy of comorbidities [3–5]. Also for ANCA-associated vasculitis real-world data is used to study epidemiology, long-term outcomes or analyze specific subgroups [6–14]. Patient identification of these studies are done base on ICD-9 or ICD-10 codes, but these codes have a notably low sensitivity and specificity for most diseases, and especially for systemic auto-immune diseases [15–17]. Recently, sensitivities from 30 to 96 % were reported for AAV specific ICD-10 codes, but with more than 30 ICD-10 codes available for vasculitis, frequently other codes are used with different codes being used between and within specialisms [17,18]. The latter is self-evident for a heterogenous, multi-organ systemic disease as AAV requiring care of several disciplines, scattering the patients and registration over different specialisms and different ICD-10 codes within health care organizations [18]. Accurate, sensitive, patient identification, however, is important to perform analyses representative of the real-world clinical practice.

Artificial intelligence (AI) based tools are increasingly used to identify relevant information within the EHR for patient identification or data collection [19–34]. The potential of AI-tools that extract features of EHRs and use Natural Language Processing (NLP) on clinical notes is also recognized by the FDA in post-market safety surveillance [35]. EHRs contain on the one hand structured data about diagnosis registration, vital signs, laboratory results and medication prescriptions, and on the other hand unstructured data like free text notes of physicians and nurses. Especially clinical notes and correspondences contain detailed information about complex diagnoses and disease courses like AAV and are often crucial to reliably establish the diagnosis. These unstructured data from clinical notes can be approached using text-mining. With a text-mining search-tool, unstructured text can be searched for the presence of keywords. However, since these keywords can also be mentioned as a consideration or a negation, text-mining will return a high false-positive rate. With NLP, a form of AI, the surrounding context of keywords can be interpreted recognizing negative (no, family history) and positive (conclusion, diagnosis) surroundings, and a score will be calculated to estimate the chance of the patient actually having the mentioned disease or complaint [30,31,36]. When an AI tool incorporates text-mining and NLP, the high false-positive rate of text-mining can be significantly reduced, increasing the accuracy [20,21,25,37].

Therefore, in the present study we hypothesized that AI can help facilitate real-world AAV research by employing text-mining and NLP on the wealth of structured and unstructured data of EHRs to accurately identify known AAV patients within large EHR-systems. To address this, we developed and assessed an identification method for AAV patients using a text-mining and NLP incorporating search tool.

2. Methods

2.1. Training center

Development of the identification method was done with the EHR-system of the Leiden University Medical Center (LUMC), an academic care center with an AAV expertise center in the Netherlands. A cohort of previously identified patients with a clinical AAV diagnosis in the LUMC (n = 226) was used to establish a training set [38]. This cohort was identified in 2019 b y manual chart review of all patients with a positive ANCA test or a registered diagnosis of AAV or a registered AAV ICD-10 code between 2013 and 2018. Because the goal is identifying AAV patients for future real-world research, we excluded patients who did not receive AAV related care, including patients who were solely referred for kidney transplantation due to AAV. Patients solely referred for kidney transplantation were excluded to guarantee applicability of the identification method in non-transplantation. Moreover, including these patients would have an inherent risk of overfitting since diagnosis of kidney failure, including AAV, is structurally reported and registered in kidney transplant registries.

2.2. Development of a text-mining search

The identification method consisting of a text-mining search and NLP-based exclusion was developed and executed using the AI-constructed CTcue Patient Finder (Ctcue B·V., version 4.5.0, Amsterdam, the Netherlands). This text-mining and NLP-based search tool was incorporated in the EHR-system of each center and searched structured and unstructured data in all patient records. An overview of the methods is shown in Fig. 1. Our aim was to identify all AAV patients and predefined the requirement that the text-mining search should achieve a sensitivity of 95 % or higher.

An iterative process guided the development of the text-mining search and was based on the training set described above. Before developing a combined search, multiple search queries were tested individually on the training set. Each query consists of selected search terms being searched in a specific data field of the EHR. Search terms were chosen based on the medical expertise of the authors. Discordances due to synonyms or frequent typing errors were noted and added to the synonym list in an iterative way. Different data fields with structured (e.g. laboratory measurements and appointments) and unstructured data (e.g. clinical notes and correspondence) were searched. For some data fields specific filters could be applied (e.g. conclusion of the clinical note or a specific time period). Then, multiple search queries that retrieved more than the predefined 95 % of the training set were combined to one overall search strategy. The final combined text-mining search was selected based on the lowest total of patients identified, while still retrieving at least 95 % of patients in the training set.

2.3. Applying text-mining search in training center

This final text-mining search was run on the complete EHR-system. As EHR-system the LUMC used Chipsoft HIX (version 6.1 HF159) which includes the records of more than 2 million patients. Only patients meeting all queries were returned by the Ctcue patient finder.

2.4. Manual review and assessment

For full assessment of performance, all identified patients were manually reviewed for a clinical AAV diagnosis by the two authors (JRL and YKOT) where discrepancies were solved by consensus. Manual review was done based on anonymized physician correspondence shown by the Ctcue interface. The study was approved by the LUMC ethics committee.

Performance was assessed by calculation of sensitivity (based on training set), positive predictive value (PPV) and the number of records needed to screen (NNS) to identify an AAV patient (1/PPV). Since only positive results were manually reviewed, specificity and negative predictive value (NPV) could not be calculated. However, when searching for a rare disease with a prevalence of less than 0.1 % within a complete EHR-system (>2.000.000), NPV and specificity can be assumed high, but are considered less or even not relevant.

2.5. NLP-based exclusion

NLP-based exclusion was added to the identification method to increase the performance. The Ctcue patient finder incorporated a preset NLP-tool, which was not optimized or adapted for this study. The tool calculates confidence scores from −100 % to +100 % by interpreting the context surrounding a search term for positive words (e.g. conclusion and history) or negative words (e.g. no and unlikely). A high negative score indicates a high chance of the patient not having the mentioned disease, whereas a high positive score indicates a high chance of the patient having the mentioned disease. Patients were excluded based on the confidence score for terms of AAV diagnoses. For different thresholds (with an interval of 10 %: −90 % to +90 %) we calculated sensitivity (based on all identified AAV patients) and PPV. For the reporting

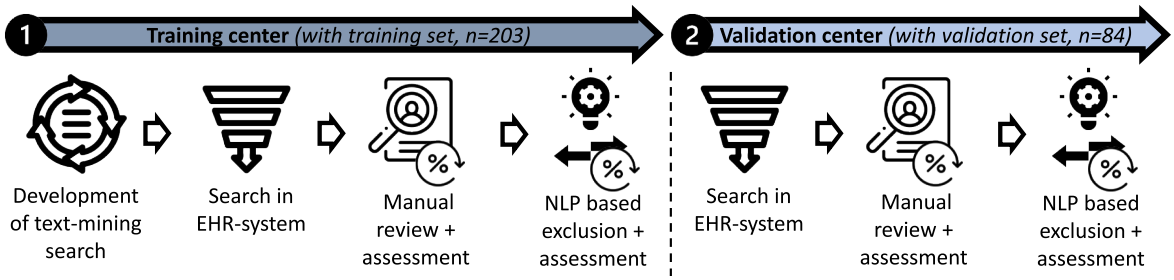

Fig. 1. Overview of the methods
EHR, Electronic healthcare records; NLP, Natural language processing.

of the final performance, we selected the highest threshold which still gave a sensitivity uphold the predefined sensitivity of 95 %.

2.6. Validation

Validation of our identification method was done in a separate hospital: the Northwest Clinics (Alkmaar, The Netherlands), a peripheral secondary care center with an expert team for AAV patients. As EHR-system Chipsoft HIX (version 6.2 HF84.1) was used. Patients previously found eligible for a prospective AAV study were used as a validation set (n = 84) [39]. These patients were identified in 2015 after screening of patients with a positive ANCA test.

In the validation center the identification method was performed in the same way as in the training center, including the text-mining search, manual review, NLP-based exclusion and assessment of performance. To verify the selected threshold for the confidence score was correct, we assessed performance of all thresholds in the same manner as in the training center.

2.7. Analyses of ICD-10 codes

To determine any selection bias when using ICD-10 codes, we compared characteristics between AAV patients with and without an AAV specific ICD-10 code (M30.1, M31.3, M31.7) in the training center. Patient, disease and treatment characteristics were analyzed by adding these queries to the Ctcue patient finder. Differences between the training set and the additionally identified AAV patients were also analyzed.

2.7.1. Statistical analysis

Sensitivity and positive predictive value are calculated with a 2×2 table and 95 % confidence intervals (95%CI) are calculated as exact Clopper-Pearson confidence intervals. Hypothesis testing for distributional differences was performed using the Mann-Whitney *U* test for numerical data and the Pearson χ^2 test for binary data with a $p < 0.05$ considered significant. All analyses were performed using R Statistical Software (v4.2.1; R Core Team 2022) and R-studio (version 2022.02.3, R Studio team 2022).

3. Results

3.1. Training set

For the training set a previously identified cohort of patients with a clinical AAV diagnosis in the LUMC (n = 226) was screened. After exclusion of 23 patients (insufficient data in the EHR (n = 1), sole treatment referral for kidney transplantation (n = 17), referral for other treatment not related to AAV diagnosis (n = 5)), 203 patients were included in the training set.

3.2. Development of a text-mining search

We summarized the iterative process of testing search queries in Table 1, depicting the search terms and data field used in each query and the number of patients identified within the complete EHR-system and within the training set. A complete list of all (partly Dutch) search terms used in the queries can be found in Supplementary Table 1.

By combining multiple queries a text-mining search with the predefined sensitivity of at least 95 % was developed. The optimal text-mining search consisted of four queries as summarized in Fig. 2A. Query 1 was based on disease description terms, combining a diagnosis term and mentioning of proteinase 3 (PR3), myeloperoxidase (MPO) or ANCA in the same physician correspondence, because the diagnosis is always specified by the presence or absence of antibodies (e.g. PR3+ GPA). This query was divided in 1 A and 1 B, because only the diagnosis terms (1 A) were later on used for calculation of confidence scores. Query 2 consisted of measurement of ANCA, PR3 or MPO in the laboratory, independent of the result. Query 3 required mentioning prednisone or cotrimoxazole in physician correspondence that also mentioned the diagnosis (1 A). For query 4 patients needed to have an appointment or admission at a specialisms involved in the treatment of AAV since 2010 (nephrology, Ear-Nose-Throat specialist (ENT), rheumatology, pulmonology, neurology, ophthalmology, dermatology, cardiology, internal medicine or pediatrics).

3.3. Applying text-mining search in training center

The final text-mining search was applied to the complete EHR-system. As depicted in Fig. 2A, each query reduced the number of identified patients considerably, with a final identification of 608 patients, including 197/203 of the training set, indicating a sensitivity of 97.0 % (197/203; 95%CI [93.7, 98.9]). Of the six patients missed by the search 3 patients had no physician correspondence at all (often related to a short follow-up period), 2 patients had no mention of prednisolone and cotrimoxazole (although they did receive the treatment) and 1 patient had no laboratory results since the primary treating physician of AAV was elsewhere and the patient only received highly-specialized ENT treatment in the LUMC.

3.4. Manual review

Manual review revealed 149 additional AAV patients, resulting in a total of 346/608 patients with a confirmed diagnosis of AAV, 238 patients with no AAV diagnosis and 24 AAV patients who met the exclusion criteria (i.e. had AAV but did not receive AAV related care including 17 patients referred for kidney transplantation). As such, the PPV was 56.9 % (346/608; 95%CI [52.9, 60.1]) and the NNS 1.76 [1.66–1.89].

3.5. NLP-based exclusion

Subsequently, the identification method was enriched with exclusion of NLP-based confidence scores resulting in a reduction of false positives

Table 1
All tested search queries for the text-mining search.

Search terms	Data field [filter]	Total patients identified	Patients (%) from training set (n = 203)
Disease description[a]			
Diagnosis (synonyms of AAV, GPA, MPA, EGPA, pauci-immune glomerulonephritis and small vessel vasculitis)	Clinical notes [conclusion]	7.608	200 (98.5)
Diagnosis	Physician correspondence	5.924	200 (98.5)
Diagnosis AND (PR3 or MPO or ANCA)	Physician correspondence	1.169	198 (97.5)
BVAS score	BVAS form	172	93 (45.8)
"BVAS" or "relapse" or "remission"	Clinical notes	190	103 (50.7)
"BVAS" or "relapse" or "remission"	Physician correspondence	26.716	148 (72.9)
Laboratory			
ANCA or PR3 or MPO measurement	Laboratory	8.449	201 (99.0)
ANCA or PR3 or MPO measurement	Laboratory [positive result]	1.365	184 (90.6)
Medication[a]			
Prednisone	Medication prescriptions	39.539	175 (86.2)
Prednisone	Clinical notes	51.168	201 (99.0)
Prednisone	Physician correspondence	53.745	197 (97.0)
Cotrimoxazole	Physician correspondence	448	146 (71.9)
Prednisone or Cotrimoxazole	Physician correspondence	60.524	198 (97.5)
Rituximab or cyclophosphamide	Physician correspondence	8387	164 (80.1)
Prednisone AND Diagnosis	Physician correspondence	1.168	197 (97.0)
(Prednisone or Cotrimoxazole) AND diagnosis	Physician correspondence	1.238	198 (97.5)
(Rituximab or Cyclophosphamide) AND diagnosis	Physician correspondence	474	165 (81.3)
Specialists			
AAV specialisms[b]	Appointment [≥1-1-2010]	304.171	200 (98.5)
AAV specialisms[b]	Appointment [≥1-1-2010] or Admission [≥1-1-2010]	457.739	202 (99.5)

ANCA, Anti-neutrophil cytoplasmatic antibody; AAV, ANCA associated vasculitis; BVAS, Birmingham vasculitis activity score; EGPA, eosinophilic granulomatosis with polyangiitis; GPA, granulomatosis with polyangiitis; MPA, microscopic polyangiitis; MPO, Myeloperoxidase; PR3, proteinase 3.

[a] A complete list of search terms and synonyms can be found in Supplementary Table 1.

[b] AAV specialisms: Nephrology, ENT, Rheumatology, Pulmonology, Neurology, Ophthalmology, Dermatology, Cardiology, Internal medicine, Pediatrics.

leading to improved PPVs as shown in Fig. 2B and detailed in Supplementary Table 3. When excluding patients with a confidence score below +80 %, 435 patients remained, including 339 of the 352 AAV patients. This results in a sensitivity of 96.3 % (339/352; 95%CI [93.8, 98.0]), which was above the predefined limit, with an increase of PPV from 56.9 to 77.9 % (339/435; 95%CI [73.7, 81.7]) and a decrease of the NNS from 1.76 to 1.28 [1.22–1.35].

3.6. Validation

We validated our identification method in a separate validation center that harbored a validation set of 84 previously identified AAV patients. The text-mining search identified 333 patients within the EHR system (Fig. 2A). This included 82/84 patients of the validation set and 112 additional identified AAV patients upon manual review, indicating a sensitivity of 97.6 % (82/84; 95%CI [91.6, 99.7]) and a PPV of 58.2 % (194/333; 95%CI [52.8, 63.6]).

When applying NLP and excluding patients with a confidence scores below +80 %, 223 patients remained, including 192 of the 196 AAV patients. As such the sensitivity remained high (98.0 % (192/196; 95% CI [94.9, 99.4])) and PPV increased from 58.2 to 86.1 % (192/223; 95% CI [80.9, 90.4]) (Fig. 2C), which was even higher than in the training center (Table 2). This also reduced the NNS from 1.72 [1.57–1.89] to 1.16 [1.11–1.24]. Detailed results of all NLP-thresholds can be find in Supplementary Table 3.

3.7. Analysis of ICD-10 codes

Of all 352 identified AAV patients in the training center, 75 % (263/352) had an AAV specific ICD-10 code registered and 25 % (89/352) did not. When comparing these groups (Table 3), in the patient group without an ICD-10 code, patients were significantly more often ANCA negative and this group contained more MPO positive patients. Furthermore, this patient group more often had fewer specialisms involved (2 vs 4), mirrored in fewer patients treated by nephrology, pulmonology, ENT, ophthalmology or neurology.

Focusing on the 149 additionally identified AAV patients, most patients visited the hospital after establishment of the training set (2018), but also more ANCA-negative AAV were identified than registered in the training set (Supplementary Table 2).

4. Discussion

In the present study we developed an AI-based identification method to identify AAV patients with a high accuracy within EHR-systems and demonstrated the added-value of NLP to increase accuracy of text-mining. The method outperformed ICD-10 registration and revealed potential biases when conducting research in patients with ICD-10 based AAV diagnoses. Thus, accurate identification with our NLP-based identification method will significantly increase sensitivity and reduce the efforts of manually reviewing patient records, thereby empowering the use of real-world data to conduct research and health care evaluation in AAV.

Unique to our study design is the fact that training and validation was done in two different hospital centers with two established sets of AAV patients. One previous study studied algorithms to identify AAV patients within EHRs and reported that combining ICD-9 codes with structured data on medication, appointments and laboratory results could improve PPV [32]. However high PPVs, up to 100 %, were at expense of the sensitivity [8–88 %], reducing the accuracy of the algorithm. Moreover, the study encompassed a high risk of overfitting bias, since sensitivity could only be determined in one training set. This training set was based on ICD-9 codes, whereas ICD-9 codes were also part of the identification algorithm, further increasing the risk of overfitting [32]. Also in previous reported identification methods of other auto-immune diseases, training sets are established based on ICD-codes and ICD-coding is often part of the studied algorithm, unintentionally causing overestimation of the sensitivity [19,22,23,25,32,33]. Additionally PPV can be projected overly positive when either the source population is narrowed to a part of the EHR-system or only a selection of identified patients is manually reviewed [19,20,22,23,25,32,33]. These studies risk an over-estimation of performance, especially for low-prevalence diseases, limiting their reliability in real-world practice. Therefore, the current study was designed to not include criteria used

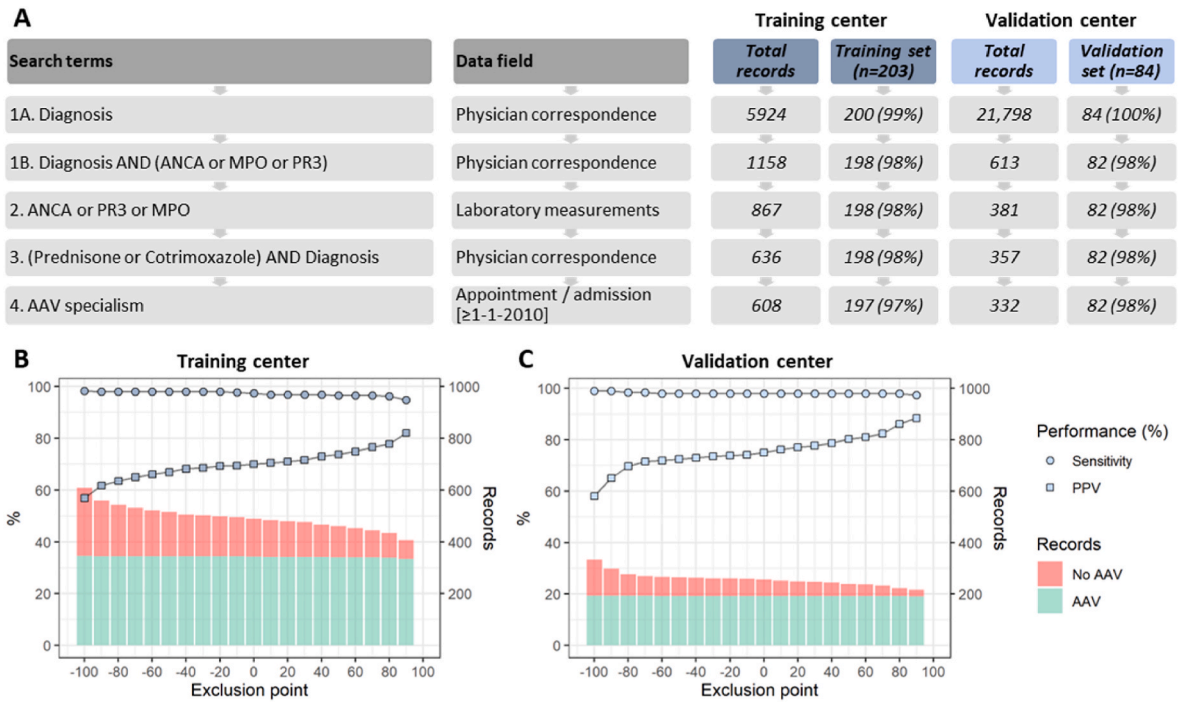

Fig. 2. Results of the text-mining search (A) and exclusion by NLP (B) in the training center.
A: The final four query search strategy. For each step the total number of patients identified in the training and validation center and the number of patients from the training and validation set retrieved are depicted.
*synonyms of MPA, GPA, EGPA, AAV, pauci-immune glomerulonephritis and small vessel vasculitis. A complete list of synonyms can be found in Supplementary Table 1.
B and C: Performance of the identification method after NLP-based exclusion of different confidence scores in the training center(B) and validation center (C). Confidence scores are based on search query 1 A. The number of found patient records with and without AAV are also depicted.
ANCA, Anti-neutrophil cytoplasmatic antibody; AAV, ANCA associated vasculitis; MPO, Myeloperoxidase; NLP, Natural language processing; PPV, positive predictive value; PR3, proteinase 3.

Table 2
Results of identification method before and after NLP in training and validation center.

	Training center		Validation center	
	Text-mining	Text-mining + NLP	Text-mining	Text-mining + NLP
Total patients identified	608	435	334	223
Sensitivity (%) [95%CI]	97.0 [93.7, 98.9][a]	96.3 [93.8, 98.0][b]	97.6 [91.6, 99.7][a]	98.0 [94.9, 99.4][b]
Positive predictive value (%)	56.9 [52.9, 60.1]	77.9 [73.7, 81.7]	58.2 [52.8, 63.6]	86.1 [80.9, 90.4]
Number needed to screen	1.76	1.28	1.72	1.16

Results of identification method in training center and validation center for the text-mining search only and text-mining combined with exclusion of NLP-based confidence scores <+80 %.
AAV, Anti-neutrophil cytoplasmatic antibody (ANCA) associated vasculitis; NLP, Natural language processing; 95%CI, 95 % confidence interval.
[a] Based on training/validation set.
[b] Based on all identified AAV patients.

during the establishment of the training and validation sets in the final identification method. By applying the identification method to two complete EHR-systems and manually reviewing all identified patients the presently reported performance of the AI-based identification method has a reliable validity that can be extrapolated to other EHRs, overcoming an often perceived barrier for implementation of AI-based tools and methods.

We refrained from using ICD-10 codes in our identification method,

Table 3
Comparison of characteristics between patients with and without an ICD-10 code registered.

	Patients with AAV ICD10 code (n = 263)	Patients without AAV ICD10 code (n = 89)	*P*-value[a]
Females (%)	99 (38)	38 (43)	0.398
Median age [IQR]	67 [54–76]	71 [58–79]	0.056
Deceased (%)	36 (14)	26 (29)	<**0.001**
ANCA positive laboratory result (%)	240 (91)	67 (75)	<**0.001**
PR3 positive	155 (59)	26 (29)	<**0.001**
MPO positive	90 (34)	43 (48)	**0.018**
Median number of treating specialisms [IQR]	4 [2–5]	2 [1–3]	<**0.001**
Treating specialisms (%)			
Nephrology	227 (86)	63 (71)	<**0.001**
ENT	176 (67)	39 (44)	<**0.001**
Cardiology	135 (51)	39 (44)	0.221
Pulmonology	134 (51)	27 (30)	<**0.001**
Ophthalmology	119 (45)	27 (30)	**0.014**
Dermatology	93 (35)	23 (26)	0.099
Neurology	91 (35)	15 (17)	**0.002**
Rheumatology	68 (26)	24 (27)	0.837
Pediatrics	6 (2)	1 (1)	0.499

ANCA, Anti-neutrophil cytoplasmatic antibody; AAV, ANCA associated vasculitis; ENT, Ear nose and throat specialist, F, female; IQR, inter quartile range; MPO, Myeloperoxidase; PR3, proteinase 3.
[a] $P < 0.05$ is considered significant.

because of the reported low sensitivity of ICD-10 codes and ICD-9 codes for AAV, especially when combined with structured data [17,32]. ICD-based algorithms are often used in claim-based studies, such as Medicare studies [40,41]. These algorithms only include patients with AAV ICD-codes registered as admission reason or patients with AAV ICD-codes registered multiple times within a time frame [6–11,40–44]. By adding criteria to ICD-10 codes, these methods can improve PPV (even up to 97 %) [7,8], but also reduce the overall sensitivity of the identification method and introduce substantial selection bias. Importantly, we were able to confirm the low sensitivity (75 %) of ICD-10 codes for AAV in the training center. Consequently, we observed ICD-10 coding led to generalized bias towards underrepresentation of ANCA negative, MPO positive and organ limited patients. Altogether, these finding corroborate the limitations of ICD-based research.

Instead of using ICD-10 codes, we started with identifying structured and unstructured data features with high sensitivity based on our training set. This included the notions that when searching for words in free text physician correspondence gave more focused results than clinical notes (likely correspondence contains less considerations and more straightforward and convincing diagnoses); combinations of words in one correspondence turned out to further exclude patient records efficiently; laboratory measurement of ANCA identified patients better than searching for a positive ANCA laboratory result (exemplary are referral patients who do not necessarily have a concurrent positive ANCA test in the EHR but might have in a previous hospital); searching for medication in correspondence worked better than searching in prescriptions (plausibly because patients could only have had immunosuppressive treatment in another previous hospital); adding admissions to appointments included patients with follow-up at another hospital after admission. As such, our text-mining search achieved our predefined requirement of a sensitivity above 95 % (i.e. 97 %) and was further implemented in this study.

The ultimate pursuit in any identification method is a 100 % performance for both sensitivity and PPV. Without NLP, text-mining returned a high false positive rate (PPV of 57–58 %) corroborating well-described limitations of text-mining [31]. Because NLP enabled interpretation of the surrounding context of diagnosis terms, the false positive rate effectively decreased, significantly increasing the PPVs from 57-58 % to 78–86 % and decreasing the records needed to screen to identify 100 AAV patients from 172 to 176 to 116–128. During validation in a second hospital with different treating physicians, performance was even better, demonstrating that the identification method does not suffer from overfitting to the training set or to physician-specific habits or manners of correspondence. By applying manual review of all patient records without any NLP confidence threshold, we were also able to verify the threshold of $<$80 % was fitting to achieve a sensitivity of $\geq$95 %. The high performance illustrates the portability of the identification method to different centers, which is an important facilitator for scaling and can significantly contribute to AAV patient identification for regional and national registries. Moreover, in accordance with statements of the FDA, the acceptability and optimal performance of AI-tools relates to its intended use and accompanying risks for patients [35]. Therefore, the currently reported performance, i.e. mean PPV of 78–86 % and sensitivity of 96–98 %, is essential to support the application of this novel search method and assess whether the method can fit the purposes of researchers and clinicians who intends to use real-world data of AAV patients and whether or not an additional manual review is required. Additionally, future studies are warranted to assess whether this identification method could also be employed for other poorly registered, rare diseases.

Some limitations of the AI-based identification method need to be addressed. Firstly, even though we determined high sensitivities in two separate cohorts, it is impossible to know exactly how many AAV patients were missed since it is impossible to know the true number of AAV patients in each hospital. Hitherto, this makes it scientifically impossible to calculate specificity or NPV, but these performance indicators have limited value when identifying rare patients in large EHR-systems. Secondly, because steroids or cotrimoxazole are included in the text-mining search, we recommend cautious monitoring whether the identification method is influenced by future treatments that completely replace steroids, requiring updating of the search terms to retain performance. Thirdly, when looking into the seven AAV patients wrongly excluded by NLP in the training center, four had a conclusion of “not active” or “limited” AAV, which due to negation caused a negative confidence score, illustrating an intrinsic limitation of NLP. When machine learning-based NLP is employed to medical data on a larger scale, accuracy of predictions will improve and, looking in to the future, NLP might even be able to assist, improve or replace registration of ICD-10 or billing codes, reducing the workload of physicians [34,36]. Another limitation of the identification method is that a specific AI tool must be installed in the hospital and integrated with the EHR-system. In this study two similar EHR-systems were used, but we can plausibly anticipate the used data fields (correspondence and laboratory measurements) are present in any available EHR-system. Therefore, we can expect that there is no limitation to applying the tool to other EHR-systems. Another benefit of this tool was the easy-to-use interface, allowing identification by physicians and researchers without any experience in ICT or artificial intelligence. When using other similar AI- tools, we can plausibly assume the text-mining search should return the comparable results since only commonly used data fields were used. However, we showed NLP has an added-value but is not perfect, emphasizing the importance of assessing the performance of NLP-tools allowing a balanced interpretation of NLP-based text-mining results. Lastly, it should be noted that ICD-10 coding was implemented in 2013, introducing a potential bias that patients treated before then were not identified by ICD-10 coding. To reassure our study’s validity, we confirmed that only one patient had no visits at the hospital after implementation of ICD-10 coding.

5. Conclusion

We describe an AI-based identification method using NLP that can accurately identify patients with an AAV diagnosis in EHRs, while avoiding bias of ICD10-coding. Therefore, this AI-based method can reduce efforts to find rare AAV patients in EHRs. The reliable high performance as well as portability of the identification method allow broad implementation, thereby fostering future AAV studies using real-world data.

Declaration of competing interest

The work of YKOT is supported by the Dutch Kidney Foundation (17OKG04) and by the Arthritis Research and Collaboration Hub (ARCH) foundation. ARCH is funded by Dutch Arthritis Foundation (ReumaNederland). YKOT received an unrestricted research grant from GlaxoSmithKline, Aurinia Pharmaceuticals and Vifor Pharma. The LUMC received consulting fees from Aurinia Pharmaceuticals, Novartis, GSK, KezarBio, Vifor Pharma, Otsuka Pharmaceuticals on consultancies delivered by YKOT.

Acknowledgements

We thank E. Dirikgil (LUMC, Leiden, the Netherlands) for providing the patient cohort used as training set and E. Houben and Y. Vegting (Northwest Clinics, Alkmaar, the Netherlands) for providing the patient cohort used as validation set. We thank L. van Heesch for access to the CTcue patient finder and D. de Jong and M. Rijnen for assistance during the use of CTcue. An abstract and poster of this work was presented at the European congress of rheumatology (EULAR) 2023 and the American Society of Nephrology (ASN) Kidney Week 2023.

This work was supported by Vifor Pharma (recipient: prof. Dr. Y.K.O. Teng), but the sponsor did not have any involvement in the study.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.compbiomed.2023.107757.

References

[1] B. Hellmich, et al., EULAR recommendations for the management of ANCA-associated vasculitis: 2022 update, Ann. Rheum. Dis. (2023). Published Online First: 16 March 2023.
[2] A.R. Kitching, et al., ANCA-associated vasculitis, Nat. Rev. Dis. Prim. 6 (1) (2020).
[3] L. Blonde, et al., Interpretation and impact of real-world clinical data for the practicing clinician, Adv. Ther. 35 (11) (2018) 1763–1774.
[4] S. de Lusignan, L. Crawford, N. Munro, Creating and using real-world evidence to answer questions about clinical effectiveness, J. Innovat. Health Inf. 22 (3) (2015) 368–373.
[5] J.M. Franklin, S. Schneeweiss, When and how can real world data analyses substitute for randomized controlled trials? Clin. Pharmacol. Ther. 102 (6) (2017) 924–933.
[6] P.M. Bataille, et al., Epidemiology of granulomatosis with polyangiitis and microscopic polyangiitis in adults in France, J. Autoimmun. 133 (2022).
[7] K.E. Nelveg-Kristensen, et al., Increasing incidence and improved survival in ANCA-associated vasculitis-a Danish nationwide study, Nephrol. Dial. Transplant. 37 (1) (2022) 63–71.
[8] L. Nygaard, et al., Long-term cardiovascular outcomes and temporal trends in patients diagnosed with ANCA-associated vasculitis: a Danish nationwide registry study, Rheumatology 62 (2) (2023) 735–746.
[9] J. Li, et al., The frequency of ANCA-associated vasculitis in a national database of hospitalized patients in China, Arthritis Res. Ther. 20 (1) (2018) 226.
[10] E. Berglin, et al., Anti-neutrophil cytoplasmic antibodies predate symptom onset of ANCA-associated vasculitis. A case-control study, J. Autoimmun. 117 (2021), 102579.
[11] M. Rivera, et al., Reasons for hospitalization and in-hospital mortality for anti-neutrophil cytoplasmic antibody vasculitides: analysis of the National Inpatient Sample, Clin. Rheumatol. 41 (1) (2022) 159–166.
[12] M. Scherlinger, et al., Worldwide trends in all-cause mortality of auto-immune systemic diseases between 2001 and 2014, Autoimmun. Rev. 19 (6) (2020).
[13] N. Droz, et al., Recurrent nephritis and/or pulmonary hemorrhage in patients with anti-glomerular basement membrane disease with and without ANCA positivity, Glomerular Dis. 1 (2) (2021) 60–67.
[14] P. Ungprasert, et al., Inpatient epidemiology and economic burden of granulomatosis with polyangiitis: a 10-year study of the national inpatient sample, Rheumatology 59 (12) (2020) 3685–3689.
[15] H. Quan, et al., Assessing validity of ICD-9-CM and ICD-10 administrative data in recording clinical conditions in a unique dually coded database, Health Serv. Res. 43 (4) (2008) 1424–1441.
[16] S. Bernatsky, T. Linehan, J.G. Hanly, The accuracy of administrative data diagnoses of systemic autoimmune rheumatic diseases, J. Rheumatol. 38 (8) (2011) 1612–1616.
[17] C. Mettler, et al., Validation of anti-neutrophil cytoplasm antibodies associated vasculitides diagnosis codes from the electronic health records of two French university hospitals, Eur. J. Intern. Med. 103 (2022) 115–117.
[18] J.B.T. Spierings, E. Dirikgil, H.B. Moens, Overdaad aan ICD-coderingen hindert onderzoek, Med. Contact 22 (2020) 18–20.
[19] A. Barnado, et al., Developing electronic health record algorithms that accurately identify patients with systemic lupus erythematosus, Arthritis Care Res. 69 (5) (2017) 687–693.
[20] T.E. Brunekreef, et al., Text mining of electronic health records can accurately identify and characterize patients with systemic lupus erythematosus, ACR Open Rheumatol 3 (2) (2021) 65–71.
[21] R.J. Carroll, et al., Portability of an algorithm to identify rheumatoid arthritis in electronic health records, J. Am. Med. Inf. Assoc. 19 (e1) (2012) e162–e169.
[22] L. Jamian, et al., Rule-based and machine learning algorithms identify patients with systemic sclerosis accurately in the electronic health record, Arthritis Res. Ther. 21 (1) (2019) 305.
[23] A. Jorge, et al., Identifying lupus patients in electronic health records: development and validation of machine learning algorithms and application of rule-based algorithms, Semin. Arthritis Rheum. 49 (1) (2019) 84–90.
[24] J. Kirshner, et al., Automated electronic health record-based tool for identification of patients with metastatic disease to facilitate clinical trial patient ascertainment, JCO Clin Cancer Inform 5 (2021) 719–727.
[25] K.P. Liao, et al., Development of phenotype algorithms using electronic medical records and incorporating natural language processing, BMJ 350 (2015) h1885.
[26] T.D. Maarseveen, et al., Machine learning electronic health record identification of patients with rheumatoid arthritis: algorithm pipeline development and validation study, JMIR Med Inform 8 (11) (2020), e23930.
[27] W.B. van Dijk, et al., Text-mining in electronic healthcare records can be used as efficient tool for screening and data collection in cardiovascular trials: a multicenter validation study, J. Clin. Epidemiol. 132 (2021) 97–105.
[28] S.A. van Laar, et al., An electronic health record text mining tool to collect real-world drug treatment outcomes: a validation study in patients with metastatic renal cell carcinoma, Clin. Pharmacol. Ther. 108 (3) (2020) 644–652.
[29] S.A. van Laar, et al., Application of electronic health record text mining: real-world tolerability, safety, and efficacy of adjuvant melanoma treatments, Cancers 14 (21) (2022).
[30] J. Wong, et al., Using machine learning to identify health outcomes from electronic health record data, Curr Epidemiol Rep 5 (4) (2018) 331–342.
[31] N. Garcelon, et al., Electronic health records for the diagnosis of rare diseases, Kidney Int. 97 (4) (2020) 676–686.
[32] A.G. Sreih, et al., Development and validation of case-finding algorithms for the identification of patients with anti-neutrophil cytoplasmic antibody-associated vasculitis in large healthcare administrative databases, Pharmacoepidemiol. Drug Saf. 25 (12) (2016) 1368–1374.
[33] S.E. Wenderfer, et al., Using a multi-institutional pediatric learning health system to identify systemic lupus erythematosus and lupus nephritis: development and validation of computable phenotypes, Clin. J. Am. Soc. Nephrol. 17 (1) (2022) 65–74.
[34] P.F. Chen, et al., Automatic ICD-10 coding and training system: deep neural network based on supervised learning, JMIR Med Inform 9 (8) (2021), e23230.
[35] Fda, Using artificial intelligence & machine learning in the development of drug & biological products [cited 2023 23-8-2023]; Available from: https://www.fda.gov/media/167973/download, 2023.
[36] N. Arivazhagan, T.T. Van Vleck, Natural Language processing basics, Clin. J. Am. Soc. Nephrol. 18 (3) (2023) 400–401.
[37] A. Esteva, et al., A guide to deep learning in healthcare, Nat. Med. 25 (1) (2019) 24–29.
[38] L.S. van Dam, et al., PR3-ANCAs predict relapses in ANCA-associated vasculitis patients after rituximab, Nephrol. Dial. Transplant. 36 (8) (2021) 1408–1417.
[39] E. Houben, et al., Prevalence and management of cardiovascular risk factors in ANCA-associated vasculitis, Rheumatology 58 (12) (2019) 2333–2335.
[40] B. Cao, et al., Polypharmacy in US Medicare beneficiaries with antineutrophil cytoplasmic antibody vasculitis, J Manag Care Spec Pharm 29 (7) (2023) 770–781.
[41] S.P. Huang, et al., Health care costs and utilization prior to diagnosis of antineutrophil cytoplasmic antibody vasculitis in Medicare beneficiaries, J Manag Care Spec Pharm 28 (11) (2022) 1292–1303.
[42] S.S. Ahn, et al., Secular trends of incidence, prevalence, and healthcare economic burden in ANCA-associated vasculitis: an analysis of the 2002-2018 South Korea national health insurance database, Front. Med. 9 (2022), 902423.
[43] B. Hellmich, et al., New insights into the epidemiology of ANCA-associated vasculitides in Germany: results from a claims data study, Rheumatology 60 (10) (2021) 4868–4873.
[44] S.T. Choi, et al., The cancer risk according to three subtypes of ANCA-associated vasculitis: a propensity score-matched analysis of a nationwide study, Semin. Arthritis Rheum. 51 (4) (2021) 692–699.