Published in final edited form as:

Semin Arthritis Rheum. 2021 February ; 51(1): 150–157. doi:10.1016/j.semarthrit.2020.10.012.

Topic Modeling to Characterize the Natural History of ANCA-Associated Vasculitis from Clinical Notes: A Proof of Concept Study

Liqin Wang, Ph.D.[1,2], **Eli Miloslavsky, M.D.**[2,3], **John H. Stone, MD, MPH**[2,3], **Hyon K. Choi, MD, DrPH**[2,3,4], **Li Zhou, Ph.D., M.D.**[1,2,*], **Zachary S. Wallace, M.D., M.Sc.**[2,3,4,*]

[1]Division of General Internal Medicine and Primary Care, Brigham and Women's Hospital, Boston, Massachusetts, USA

[2]Harvard Medical School, Boston, Massachusetts, USA

[3]Rheumatology Unit, Division of Rheumatology, Allergy, and Immunology, Massachusetts General Hospital, All in Boston, Massachusetts, USA

[4]Clinical Epidemiology Program, Mongan Institute, Massachusetts General Hospital, All in Boston, Massachusetts, USA

Abstract

Objectives—Clinical notes from electronic health records (EHR) are important to characterize the natural history, comorbidities, and complications of ANCA-associated vasculitis (AAV) because these details may not be captured by claims and structured data. However, labor-intensive chart review is often required to extract information from notes. We hypothesized that machine learning can automatically discover clinically-relevant themes across longitudinal notes to study AAV.

Methods—This retrospective study included prevalent PR3- or MPO-ANCA+ AAV cases managed within the Mass General Brigham integrated health care system with providers' notes

Corresponding Author: Zachary S. Wallace, MD, MSc, Clinical Epidemiology Program, Division of Rheumatology, Allergy, and Immunology, Massachusetts General Hospital, 100 Cambridge Street, 16th Floor, Boston, MA 02114, T: 617-724-2507; F: 617-643-1274, zswallace@mgh.harvard.edu, @zach_wallace_md.

[*]LZ and ZSW contributed equally to this manuscript

AUTHOR CONTRIBUTIONS

Drs Wang and Wallace had full access to all the data in the study and take responsibility for the integrity of the data and the accuracy of the data analysis.

Study conception and design: Wang, Choi, Zhou, Wallace

Acquisition of data: Wang, Wallace.

Analysis and interpretation of data: Wang, Miloslavsky, Stone, Choi, Wallace

Drafting of the manuscript: Wang, Wallace.

Critical revision of the manuscript for important intellectual content: All authors.

Final approval of the version of the article to be published: All authors.

CONFLICT OF INTEREST STATEMENT None declared.

Code availability: The source code for the topic stabilization and summarization of the topics over the study course has been released to the public on Github (https://github.com/bylinn/aav_topic_modeling).

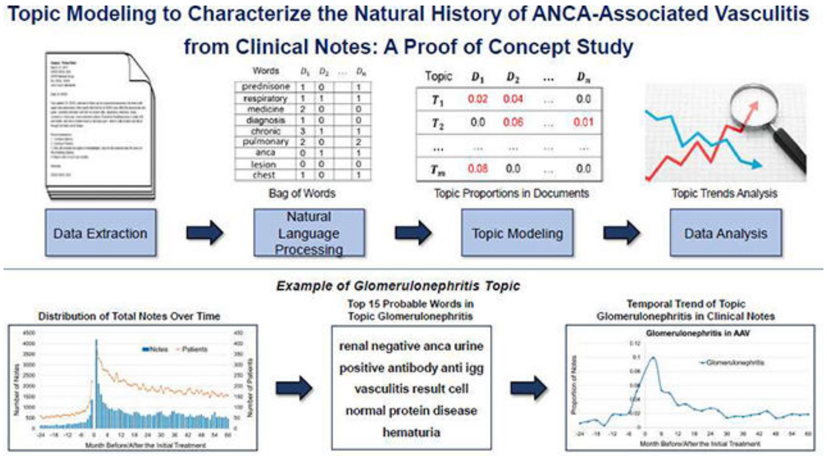
Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

available between March 1, 1990 and August 23, 2018. We generated clinically-relevant topics mentioned in notes using latent Dirichlet allocation-based topic modeling and conducted trend analyses of those topics over the 2 years prior to and 5 years after the initiation of AAV-specific treatment.

Results—The study cohort included 660 patients with AAV. We generated 90 topics using 113,048 available notes. Topics were related to the AAV diagnosis, treatment, symptoms and manifestations (e.g., glomerulonephritis), and complications (e.g., end-stage renal disease, infection). AAV-related symptoms and psychiatric symptoms were mentioned months before treatment initiation. Topics related to pulmonary and renal diseases, diabetes, and infections were common during the disease course but followed distinct temporal patterns.

Conclusions—Automated topic modeling can be used to discover clinically-relevant themes and temporal patterns related to the diagnosis, treatment, comorbidities, and complications of AAV from EHR notes. Future research might compare the temporal patterns in a non-AAV cohort and leverage clinical notes to identify possible AAV cases prospectively.

Graphical Abstract


Topic Modeling to Characterize the Natural History of ANCA-Associated Vasculitis from Clinical Notes: A Proof of Concept Study

Words	D_1	D_2	…	D_n
prednisone	1	0		1
respiratory	1	1		1
medicine	2	0		0
diagnosis	1	0		0
chronic	3	1		1
pulmonary	2	0		2
anca	0	1		1
lesion	0	0		0
chest	1	0		1

Topic	D_1	D_2	…	D_n
T_1	0.02	0.04	…	0.0
T_2	0.0	0.06	…	0.01
…	…	…	…	…
T_m	0.08	0.0	…	0.0

renal negative anca urine positive antibody anti igg vasculitis result cell normal protein disease hematuria

Bag of Words
Topic Proportions in Documents
Topic Trends Analysis
Data Extraction
Natural Language Processing
Topic Modeling
Data Analysis
Example of Glomerulonephritis Topic
Distribution of Total Notes Over Time
Notes
Patients
Number of Notes
Number of Patients
Month Before/After the Initial Treatment
Top 15 Probable Words in Topic Glomerulonephritis
Temporal Trend of Topic Glomerulonephritis in Clinical Notes
Glomerulonephritis in AAV
Glomerulonephritis
Proportion of Notes
Month Before/After the Initial Treatment

Keywords

ANCA-Associated Vasculitis; Electronic Health Records; Natural Language Processing; Topic Modeling; Epidemiology

INTRODUCTION

Anti-neutrophil cytoplasmic antibody (ANCA)-associated vasculitis (AAV) is a small to medium vessel vasculitis with manifestations that can vary widely both in their anatomic distribution (e.g., sinus, pulmonary, renal) as well as their severity (e.g., nasal crusting, rapidly progressive glomerulonephritis) (1). AAV is often associated with end-organ damage (e.g., end-stage renal disease) (2), complications of treatment (e.g., diabetes, infection, malignancy) (2–4), reduced quality of life (5), and a 2-fold higher risk of death (6). Understanding the clinical course of AAV, including the temporal evolution of clinical manifestations, comorbidities, and complications, can identify opportunities for interventions that can improve morbidity and mortality in AAV.

To date, retrospective AAV studies have been limited to case series and claims-based studies which often rely on chart review, coded fields in electronic health records (EHR), and/or billing codes (7–9). These approaches, however, have notable limitations. Chart review is time-intensive while analyses using coded fields can be limited because investigators have an incomplete or biased understanding of which data elements will be common and relevant to extract. Moreover, many of the symptoms (e.g., cough, anxiety, numbness) reported by AAV patients are often not captured in claims data (e.g., billing codes) or structured data fields (e.g., laboratory data, medications, problem lists) in the EHR. Thus, care provider notes may be a potentially rich source of data for understanding the clinical course of AAV as they cover a variety of elements, including medical history, physical exam, laboratory tests, diagnoses, medical reasoning, and care plans (10). No previous study has evaluated notes as a data source to characterize AAV's natural history.

Free-text notes, however, simply contain words and symbols and cannot be used "as-is" for studying the natural history of AAV. Topic modeling, a natural language processing (NLP) approach, applies statistical machine learning methods to identify abstract topics or themes that occur in a properly prepared collection of documents (10–15). It is an unsupervised approach that does not require any prior labeling of the documents but automatically discovers and annotates large dataset with latent "topics". In topic models, each document is modeled as a finite mixture of topics with each topic viewed as a probability distribution over a fixed vocabulary (often consists of thousands or millions of words) in the documents. Once the topics are identified, the distribution of the topics in each document can be determined and the analyses can be transferred from words to meaningful topics. Topic modeling has been widely used in processing various types of data like biomedical literature, social media, and healthcare data for different purposes, such as information retrieval, document summarization and visualization, temporal patterns discovery, and outcome predictions (16–19). In the present study, we hypothesized that we could leverage a unique data source (i.e., inpatient and outpatient provider notes) and topic modeling to identify clinically-relevant topics or themes and characterize the temporal clinical course of AAV.

METHODS

Study Cohort

The Mass General Brigham (formerly Partners) AAV cohort is a retrospective cohort established at Mass General Brigham (MGB, formerly Partners HealthCare System), a large healthcare system that includes tertiary care and community hospitals as well as primary care and specialty outpatient clinics, providing healthcare services for more than half of the greater Boston area population. All included AAV cases are proteinase 3 (PR3)- or myeloperoxidase (MPO)−ANCA+ and identified using a combination of ANCA test results, billing codes, and pre-defined keyword references in ambulatory notes (20). The date of treatment initiation for AAV was determined by chart review (20). For this study, we included prevalent cases with at least one clinical note available in the MGB databases. This study was approved by the MGB Institutional Review Board.

Data Preparation

Clinical data were collected from the MGB Research Patient Data Registry (RPDR), a centralized data warehouse for the medical records associated with care provided at MGB facilities (21, 22). We obtained inpatient and outpatient clinical notes, including ambulatory notes, admission notes, progress notes, and discharge summaries, documented between March 1, 1990 and August 23, 2018. We preprocessed the notes in two steps. First, we removed word tokens which were: (a) non-alphabetic words, (b) commonly used words, such as 'the', 'in', 'of', (c) words occurring in fewer than 25 documents, many of which were attributed to misspellings or noise, and (d) words that were human names except those being part of the *disease or finding* concept names found in a comprehensive medical terminology, the SNOMED CT (e.g., 'Wegener', 'Churg') (23). Second, we used a NLP tool kit (NLTK) (24) to stem noun phrases and verbs. After the above data preparation step, the corpus was transformed into a collection of "bag-of-words" in which the order and the local structure of the contents of the documents were ignored, which could then be used for topic modeling.

Topic Modeling, Generation and Annotation

We performed automatic topic modeling using Mallet (25), an implementation of latent Dirichlet allocation (LDA)-based topic modeling, over the entire processed corpus to generate latent topics. Due to potential variation in identified topics each time LDA is applied to the study corpus, we only included topics which occurred stably over three LDA model iterations. The identification of stable topics was based on a greedy algorithm search over the space of triplets generated by combination of topics from the three model iterations, for which the max cosine distance of any two topics in the triplets should be smaller than an adjustable parameter, or 0.7 in this study (10, 15). From the MGB AAV cohort, we retrieved a total of 113,048 notes. We used all available notes to generate three topic models with each model having 100 topics. After stabilization, we identified 90 topics. We provided the top 25 probable words of the topics for manually review. Each topic was labeled independently by two vasculitis experts (E.M., Z.W.) primarily based on the first few words and any differences were resolved through consensus. We assessed the clinical relevance of each topic to AAV including its diagnosis, manifestations, complications, and treatment. For example, a topic with the top 25 probable words of '*anca vasculitis rituximab renal skin prednisone ckd rituxan azathioprine wegener chest steroid chronic clear disease month recurrent edema improve hypertension stable past fibrosis rash anemia*' was labeled as *AAV*. Among these 90 topics, some were labeled with the same topic name because they were thought to capture a similar theme despite some differences in probable words. For example, '*bronchiectasis culture hemoptysis specimen sputum name chest neg neb negative lobe pneumonia pulmonary stable result anca lower lung infection name bronchial pseudomonas flare upper final'* and '*bronchiectasis neg pseudomonas sputum flare oral nasal mucoid chest neb hyperlipidemia inhaler disease sen stable smear post lobe heart appt increase name infection flora lung*' were labeled as *Bronchiectasis*. Furthermore, we grouped many topics into clinically relevant, more general categories; some selected topics and general categories are included in Table 2 with the top 15 of the 25 probable words included. The entire list of topics and full list of the top 25 probable words associated with each is available in Supplementary Table 1.

Statistical Analysis

We studied the temporal trends in the documentation of stable topics before and after the date that treatment was initiated for a diagnosis of AAV. We aligned the longitudinal clinical notes of the study cohort based on the distance from the date of initial treatment. For example, notes documented within the 3 months following treatment initiation were labeled as ‘3’ while the notes documented within the 3 months preceding the initial treatment were labeled as ‘−3’. We labeled each note documented within the 2 years before and 5 years after the initial treatment (based on median follow-up time before and after treatment initiation) and grouped them into 3-month intervals, resulting in a total of 28 smaller corpora (i.e., 8 before and 20 after the initial treatment), one corpus for each 3-month period. Then, for each corpus, we determined the proportion score of each topic by dividing the number of words assigned to a given topic by the total number of words in the corpus (10). These topic proportion scores were used to plot the trends of the topics over the temporal course of AAV.

To evaluate the validity of the trends generated using topic modeling of clinical notes, we selected two topics that were likely to also be reflected in structured EHR data and compared the trends generated from the two different sources. Specifically, we selected two LDA topics with one related to a specific diagnosis (i.e., *glomerulonephritis*) and another related to medication prescription (i.e., *AAV treatment*). Using data from encounter-associated diagnoses and procedure codes (for infusion medications) and medication prescriptions (for oral medications), respectively, we evaluated the percentage of patients with a diagnosis of glomerulonephritis and the percentage of patients receiving AAV treatment before and after the AAV initial treatment (Supplementary Table 2). We then compared the changes in the percentage of patients versus the proportion of notes over time and measured the correlation between two trends using Pearson’s r correlation (Supplementary Figure 1).

We used Microsoft® Excel 2016 to draw the trends of the topics based on clinically meaningful groups as assigned by E.M. and Z.W. (e.g., AAV diagnosis, AAV treatment, pulmonary involvement in AAV) and the trends from structured data. Statistical analyses were performed in R software, version 3.5.3 (R Foundation for Statistical Computing).

RESULTS

This analysis included 660 prevalent ANCA+ AAV cases with clinical notes available between March 1, 1990 and August 23, 2018. The mean age at treatment initiation was 56.9 (18.1) years, 392 (59.4%) were female, 268 (41.6%) were PR3−ANCA+, and 392 (59.4%) were MPO−ANCA+ (Table 1). The median follow-up time from the first available to the last available note was 118 (IQR 61, 180) months.

Distribution of Clinical Notes During the Study Period

Of the retrieved notes, 53,921 (40.5%) were documented during the relevant study period (2 years before to 5 years after treatment initiation). Figure 1 shows the distribution of the number of clinical notes and the number of patients with available notes by month relative to the date of treatment initiation. The median number of notes available per patient per month during the study period was 2 (IQR: 1, 4), but the greatest median number of notes per patient was during the first month after treatment initiation (7 notes/patient, IQR: 3,14).

Temporal Trends of Topics

We examined the temporal trends of these topics before and after AAV treatment initiation. The temporal trends of selected topics related to the diagnosis and treatment of AAV are shown in Figure 2. The proportion of notes referencing the diagnosis of AAV, including granulomatosis with polyangiitis (i.e., Wegener's granulomatosis) and eosinophilic granulomatosis with polyangiitis (i.e., Churg-Strauss syndrome), began to increase approximately one month before the initiation of AAV treatment and continued to represent a significant proportion of documents in the years following the diagnosis (Figure 2a). Figure 2b shows the temporal trends in topics related to the treatment of AAV. References to specific AAV treatments (e.g., rituximab, cyclophosphamide) became frequent, and continued to be frequent, following the date of treatment initiation which had been previously determined by manual chart review. In the months preceding the increase in references to AAV-specific treatment, there were also references to treatment in general, particularly glucocorticoids.

Figures 3a, 3b, and 3c include the temporal trends in references to head and neck, pulmonary, and renal manifestations. Whereas references to head and neck manifestations, pulmonary findings, pulmonary symptoms, and bronchiectasis remained relatively stable before and after the initiation of AAV treatment, references to glomerulonephritis spiked around the initiation of AAV treatment and subsequently became less common. The frequency of references to end-stage renal disease and renal transplantation increased following AAV treatment initiation. Figure 3d includes trends in reference to other potential AAV manifestations such as articular, cutaneous, and neurological symptoms. Reference to these symptoms were present months prior to the initiation of AAV treatment and continued to be referenced during the follow-up period.

Several of the identified topics included comorbidities or complications of AAV. For instance, Figures 4a and 4b show the frequency with which psychiatric symptoms and diabetes, respectively, were referenced during the study period which was similar to the frequency of selected AAV manifestations presented in Figure 3. The temporal trend of a topic related to infections followed a rather distinct pattern compared to other comorbidities and complications. Its frequency was greatest around the time of treatment initiation and in the months following this highly immunosuppressive time period but then gradually decreased in frequency over subsequent months and years. However, references to infection remained notably common in the years following treatment initiation. To further evaluate factors that may be associated with infection or suspicion for infection, we overlayed trends for topics related to pulmonary disease and head and neck manifestations with the trend for the infection topic (Supplementary Figure 2). These trends suggest that pulmonary symptoms may contribute to infectious concerns as captured in notes around the time of diagnosis.

Supplementary Figure 1a and e1b show the comparison of the trends from clinical notes based on topic modeling with those of their respective structured data elements, including diagnoses (*glomerulonephritis*) and medications (*AAV treatment*). As shown in the Supplementary Figure 1a, the trend of topic *AAV treatment* is similar to the trend based on the medication prescription, although the rate of increase/decrease is slightly different. In the Supplementary Figure 1b, the temporal pattern of the topic *glomerulonephritis* is similar to the temporal patterns based on the ICD codes. The Pearson's r correlation was 0.80 for *AAV treatment*, and 0.92 for *glomerulonephritis*, both indicating strong positive correlation.

DISCUSSION

In this proof-of-concept study, automated topic modeling was used to identify 90 clinically-relevant topics mentioned in the clinical notes of AAV patients. The temporal trends of references to these topics before and after AAV treatment initiation reflect what we understand to be the natural history of AAV, including the diagnosis, its treatments, its specific manifestations (e.g., renal disease, pulmonary manifestations), and its complications (e.g., infection). In addition to identifying topics hypothesized *a priori* to be common (e.g., diagnosis, treatment, renal disease), this method also identified references to psychiatric symptoms (e.g., depression and anxiety) which were unexpected. These findings demonstrate that automated topic modeling is feasible in AAV research and can likely be applied in other multi-organ rheumatic conditions.

We observed a strong correlation between the trends of the topics (i.e., *glomerulonephritis* and *AAV treatment*) identified from clinical notes with the trends identified using corresponding structured EHR data. This supports the feasibility of using our approach to identify topics and trends that may not be available in structured EHR data. The ability to identify clinically-relevant topics using provider notes is important in multi-organ conditions like AAV, systemic sclerosis, and systemic lupus erythematosus where structured data fields, such as billing codes, are unlikely to reflect the full extent of signs, symptoms, comorbidities, and complications associated with the diagnosis. Moreover, given the fragmentation of healthcare in the United States, where an individual's insurance coverage can change year-to-year (limiting the use of claims data) (26) and where an individual's providers might be spread out across several healthcare systems with potentially disparate encoded data (limiting the use of EHR structured data) (27–30), the ability to leverage clinical notes provides a novel opportunity to study the natural history of disease in patient cohorts, regardless of insurance coverage or where their other providers (e.g., primary care provider, pulmonologist) might be located.

In addition to establishing the feasibility of applying this method in multi-system rheumatic conditions like AAV, this study also provides insights into how aspects of the natural history of AAV can be gleaned from clinical notes. First, the observed temporal trends in various topics highlight the chronicity of AAV and its care, as well as the increased frequency of references in notes during active disease (e.g., treatment initiation). Future studies might evaluate how the frequency of references to certain topics in notes might identify other periods of increased disease activity (e.g., flare). Second, delays in diagnosis are common (31, 32) and were suggested by the frequency with which signs or symptoms related to AAV manifestations and non-specific treatments (e.g., glucocorticoids) were mentioned in the months preceding AAV-specific treatment initiation. Similarly, notes often capture a provider's clinical reasoning and we found that the diagnosis of AAV is often mentioned in clinical notes in the few months preceding AAV-specific treatment initiation. This likely reflects the diagnostic considerations made when patients present with symptoms potentially related to AAV, before diagnostic confirmation.

A strength of automated topic modeling is that it is an unsupervised method that can identify themes not expected to be present *a priori* in addition to those expected based on prior experience or knowledge. In this study, psychiatric symptoms (e.g., anxiety, depression) were identified as a common topic. Such symptoms are known to significantly impact the lives of patients with AAV (5, 33) but may be poorly documented or treated. As with signs and symptoms related to AAV manifestations, psychiatric symptoms were frequently mentioned before treatment initiation and likely reflect the anxiety and stress associated with unexplained symptoms and/or a diagnosis of a rare and potentially life-threatening diagnosis (34).

The temporal trend of glomerulonephritis contrasted significantly with that of other AAV manifestations (e.g., pulmonary disease, joint symptoms). This likely reflects the chronicity of certain symptoms (e.g., sinusitis, lung disease, neurologic) which are more indolent and may have variable responses to treatment but significantly impact a patient's daily activities and function. In contrast, acute glomerulonephritis is an emergency and, once recognized, often leads to medical encounters, the diagnosis of AAV, and prompt treatment initiation. These differences highlight the potential use of topic modeling to not only identify relevant themes but also characterize their relative importance during the clinical course of disease.

References to infectious symptoms or conditions also followed a distinct course. There were frequent mentions of this topic around the time of treatment initiation, likely reflecting the complications associated with immunosuppression (4) but also the diagnostic uncertainty confronting providers when patients present with symptoms (e.g., pulmonary symptoms) ultimately attributed to AAV. As expected, the topic is mentioned less frequently over subsequent months consistent with the fact that immunosuppression typically decreases over time and there is presumably less diagnostic uncertainty. Despite this trend, *infections* remains a frequently mentioned topic in the months and years following treatment initiation.

The strengths of this study include its application of a novel, unbiased methodology to identify common themes in AAV clinical care from a large collection of free-text notes and the use of a large AAV cohort with relatively long follow-up. Despite these strengths, this study has certain limitations. First, the study was conducted using clinical notes with a somewhat uneven distribution, such that the vast majority of notes were available from the time of and following treatment initiation with fewer notes (of fewer patients) available in the months and years preceding treatment initiation. While this may limit the ability to comment on thematic trends prior to treatment initiation, trends observed following treatment initiation can be interpreted with confidence. Second, themes identified in clinical notes may reflect the use of copy and paste functions in clinical documentation (35). However, the distinct trends observed (e.g., *glomerulonephritis*), suggest that copy-and-paste practices were unlikely to influence the observed trends in acute diseases. Third, this study was conducted from the perspective of a healthcare system so outside clinical documentation was not captured. However, the MGB includes two large tertiary care centers, community hospitals, and primary care clinics which provide many healthcare services to the population in the New England region, likely minimizing the incompleteness of notes. Fourth, potential population diversity and clinical documentation variations (e.g., using templates) among different healthcare systems suggest that the topics and trends captured from the data in one EMR system may require additional validation using data from other systems or EHRs before reaching a generalizable conclusion. However, the methodology presented in the study could be easily adaptable to other EHR data.

In summary, we found that automated topic modeling using clinical documentation identified topics relevant to the diagnosis, treatment, comorbidities, and complications of AAV. The temporal trends of these topics reflect previously described diagnostic delays and suggest that some symptoms (e.g., acute glomerulonephritis) may affect patients over the duration of their illness differently than others (e.g., pulmonary symptoms). Therefore, this method can provide unique insights regarding the clinical course of AAV that may not be captured in structured EHR data fields and claims data. Future studies might compare the temporal patterns in a non-AAV cohort and evaluate the role of the topics as features in deep machine learning models to identify possible AAV cases both retrospectively and prospectively.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

ACKNOWLEDGEMENTS

The authors thank Serena Fu and Tyler Harkness for data cleaning and administrative assistance with this research.

FUNDING

LW and LZ were supported with funding from Agency for HealthCare Research and Quality grant R01HS025375, NIH/NIAID R01AI150295, Partners Innovation Fund. ZSW supported by NIH/NIAMS [K23AR073334 and L30 AR070520 to ZSW]

REFERENCES

1. Mahr A, Katsahian S, Varet H, Guillevin L, Hagen EC, Hoglund P, et al. Revisiting the classification of clinical phenotypes of anti-neutrophil cytoplasmic antibody-associated vasculitis: a cluster analysis. Ann Rheum Dis 2013;72(6):1003–10. [PubMed: 22962314]

2. Robson J, Doll H, Suppiah R, Flossmann O, Harper L, Hoglund P, et al. Damage in the ANCA-associated vasculitides: long-term data from the European vasculitis study group (EUVAS) therapeutic trials. Ann Rheum Dis 2015;74(1):177–84. [PubMed: 24243925]

3. van Daalen EE, Rizzo R, Kronbichler A, Wolterbeek R, Bruijn JA, Jayne DR, et al. Effect of rituximab on malignancy risk in patients with ANCA-associated vasculitis. Ann Rheum Dis 2016;76(6):1064–69. [PubMed: 27899372]

4. Lafarge A, Joseph A, Pagnoux C, Puechal X, Cohen P, Samson M, et al. Predictive Factors of Severe Infections in Patients With Systemic Necrotizing Vasculitides: Data From 733 Patients Enrolled in Five Randomized Controlled Trials of the French Vasculitis Study Group. Rheumatology (Oxford). 2019;[Epub online ahead of print].

5. Basu N, McClean A, Harper L, Amft EN, Dhaun N, Luqmani RA, et al. The characterisation and determinants of quality of life in ANCA associated vasculitis. Ann Rheum Dis 2014;73(1):207–11. [PubMed: 23355077]
6. Tan JA, Dehghan N, Chen W, Xie H, Esdaile JM, Avina-Zubieta JA. Mortality in ANCA-associated vasculitis: a meta-analysis of observational studies. Ann Rheum Dis 2017;76(9):1566–74. [PubMed: 28468793]
7. Solans-Laque R, Fraile G, Rodriguez-Carballeira M, Caminal L, Castillo MJ, Martinez-Valle F, et al. Clinical characteristics and outcome of Spanish patients with ANCA-associated vasculitides: Impact of the vasculitis type, ANCA specificity, and treatment on mortality and morbidity. Medicine (Baltimore). 2017;96(8):e6083. [PubMed: 28225490]
8. Panupattanapong S, Stwalley DL, White AJ, Olsen MA, French AR, Hartman ME. Epidemiology and Outcomes of Granulomatosis with Polyangiitis (GPA) in Pediatric and Working-age Adults Populations in the United States: Analysis of a Large National Claims Database. Arthritis Rheumatol 2018;70(12):2067–76. [PubMed: 29806148]
9. Tan JA, Choi HK, Xie H, Sayre EC, Esdaile JM, Avina-Zubieta JA. All-Cause and Cause-Specific Mortality in Patients With Granulomatosis With Polyangiitis: A Population-Based Study. Arthritis Care Res (Hoboken). 2019;71(1):155–63. [PubMed: 29692001]
10. Wang L, Lakin J, Riley C, Korach Z, Frain LN, Zhou L. Disease Trajectories and End-of-Life Care for Dementias: Latent Topic Modeling and Trend Analysis Using Clinical Notes. AMIA Annu Symp Proc 2018;2018:1056–65. [PubMed: 30815148]
11. Blei DM, Ng AY, Jordan MI. Latent dirichlet allocation. Journal of machine Learning research. 2003;3(Jan):993–1022.
12. Tang C, Zhu Y, Blackley SV, Plasek JM, Wan M, Zhou L, et al. Visualizing Literature Review Theme Evolution on Timeline Maps: Comparison Across Disciplines. IEEE Access. 2019;7:90597–607.
13. Wang L, Sha L, Lakin JR, Bynum J, Bates DW, Hong P, et al. Development and Validation of a Deep Learning Algorithm for Mortality Prediction in Selecting Patients With Dementia for Earlier Palliative Care Interventions. JAMA Netw Open. 2019;2(7):e196972. [PubMed: 31298717]
14. Steyvers M, Griffiths T. Probabilistic topic models. Handbook of latent semantic analysis. 2007;427(7):424–40.
15. Shao Y, Mohanty AF, Ahmed A, Weir CR, Bray BE, Shah RU, et al., editors. Identification and use of frailty indicators from text to examine associations with clinical outcomes among patients with heart failure AMIA Annual Symposium Proceedings; 2016: American Medical Informatics Association.
16. Nagwani NK. Summarizing large text collection using topic modeling and clustering based on MapReduce framework. Journal of big data. 2015;2(1):1–18.
17. Liu L, Tang L, Dong W, Yao S, Zhou W. An overview of topic modeling and its current applications in bioinformatics. SpringerPlus. 2016;5(1):1–22. [PubMed: 26759740]
18. Griffiths TL, Steyvers M. Finding scientific topics. Proceedings of the National Academy of Sciences - PNAS. 2004;101(Supplement 1):5228–35.
19. Boyd-Graber J, Hu Y, Mimno D. Applications of Topic Models. Foundations and trends in information retrieval. 2017;11(2–3):143–296.
20. Wright A, Chen ES, Maloney FL. An automated technique for identifying associations between medications, laboratory results and problems. Journal of biomedical informatics. 2010;43(6):891–901. [PubMed: 20884377]
21. Murphy SN, Morgan MM, Barnett GO, Chueh HC. Optimizing Healthcare Research Data Warehouse Design Through Past COSTAR Query Analysis. Proc AMIA Symp. 1999:892–6. [PubMed: 10566489]
22. Murphy SN, Chueh HC. A Security Architecture for Query Tools Used to Access Large Biomedical Databases. Proc AMIA Symp. 2002:552–6. [PubMed: 12463885]
23. Donnelly K SNOMED-CT: The advanced terminology and coding system for eHealth. Studies in health technology and informatics. 2006;121:279. [PubMed: 17095826]
24. Bird S, Loper E, editors. NLTK: the natural language toolkit. Proceedings of the ACL 2004 on Interactive poster and demonstration sessions; 2004: Association for Computational Linguistics.

25. McCallum AK. Mallet: A machine learning for language toolkit. 2002.

26. Barnett ML, Song Z, Rose S, Bitton A, Chernew ME, Landon BE. Insurance Transitions and Changes in Physician and Emergency Department Utilization: An Observational Study. J Gen Intern Med 2017;32(10):1146–55. [PubMed: 28523475]

27. Bourgeois FC, Olson KL, Mandl KD. Patients treated at multiple acute health care facilities: quantifying information fragmentation. Archives of internal medicine. 2010;170(22):1989–95. [PubMed: 21149756]

28. Smith PC, Araya-Guerra R, Bublitz C, Parnes B, Dickinson LM, Van Vorst R, et al. Missing clinical information during primary care visits. Jama. 2005;293(5):565–71. [PubMed: 15687311]

29. Kern LM, Grinspan Z, Shapiro JS, Kaushal R. Patients' Use of Multiple Hospitals in a Major US City: Implications for Population Management. Popul Health Manag 2017;20(2):99–102. [PubMed: 27268133]

30. Vest JR, Gamm LD. Health information exchange: persistent challenges and new strategies. J Am Med Inform Assoc 2010;17(3):288–94. [PubMed: 20442146]

31. Poulton CJ, Nachman P, Hu Y, McGregor JA, Jennette CE, Falk RJ, et al. Pathways to renal biopsy and diagnosis among patients with ANCA small-vessel vasculitis. Clin Exp Rheumatol 2013;31(Suppl. 75):S32–S7. [PubMed: 23343774]

32. Abdou NI, Kullman GJ, Hoffman GS, Sharp GC, Specks U, McDonald T, et al. Wegener's granulomatosis: survey of 701 patients in North America. Changes in outcome in the 1990s. The Journal of rheumatology. 2002;29(2):309–16. [PubMed: 11838848]

33. Yun JD, Ha J, Kim S, Park HA, Yoo J, Ahn SS, et al. Predictor of Depressive Disorders in Patients With Antineutrophil Cytoplasmic Antibody-Associated Vasculitis. Clin Rheumatol 2019;38(12):3485–91. [PubMed: 31312987]

34. Robson JC, Dawson J, Doll H, Cronholm PF, Milman N, Kellom K, et al. Validation of the ANCA-associated vasculitis patient-reported outcomes (AAV-PRO) questionnaire. Ann Rheum Dis 2018;77(8):1157–64. [PubMed: 29695498]

35. Cohen R, Aviram I, Elhadad M, Elhadad N. Redundancy-aware topic modeling for patient record notes. PLoS One. 2014;9(2):e87555. [PubMed: 24551060]

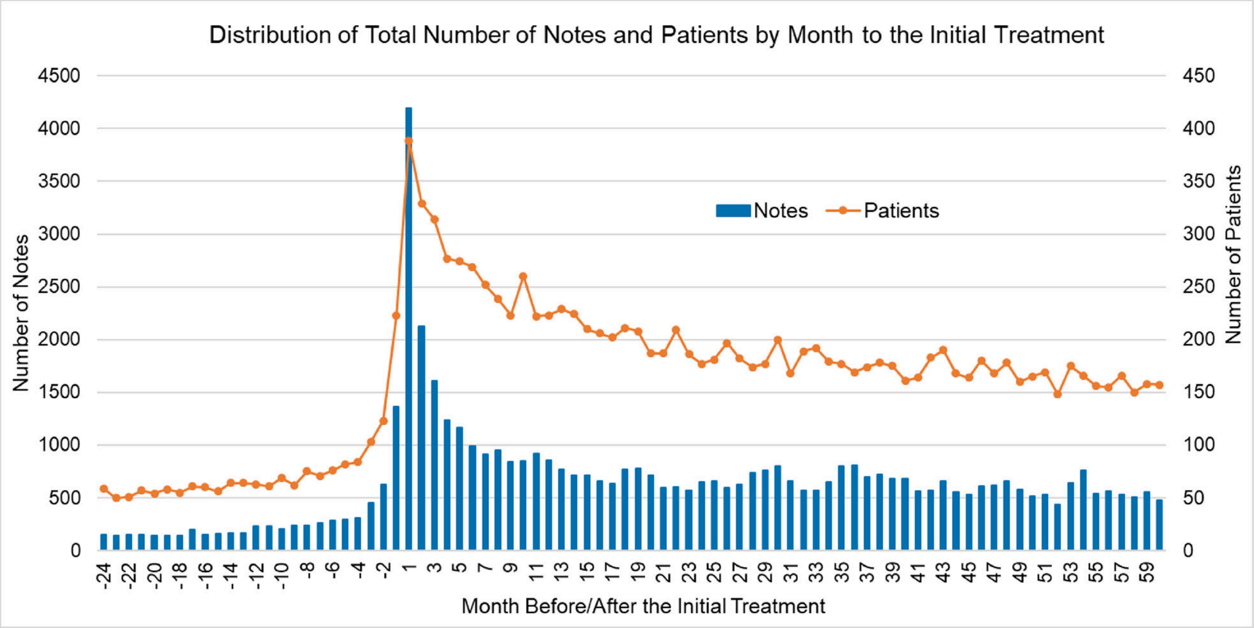


Figure 1.

The distribution of the number of clinical notes and patients by month relative to the date receiving initial treatment.

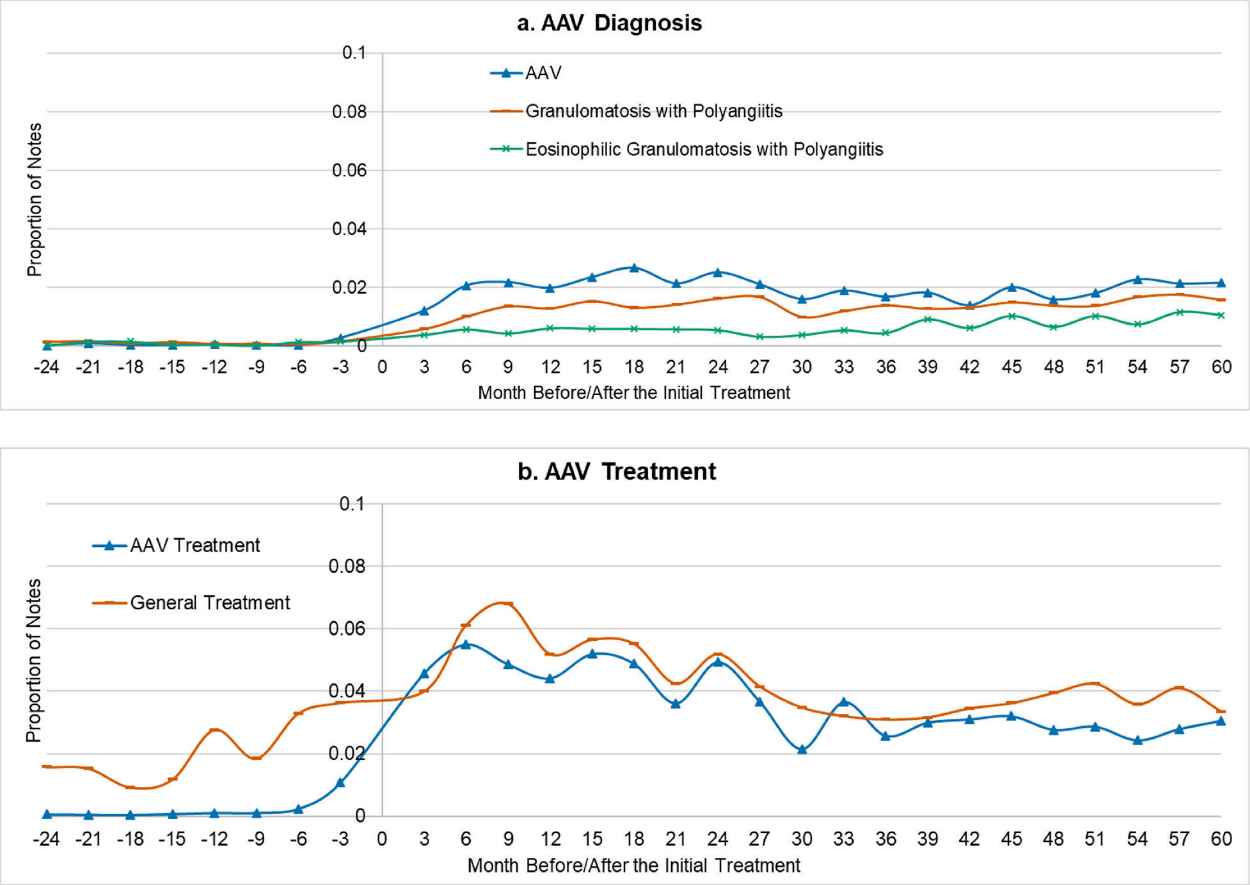


Figure 2.

AAV diagnosis and treatment. A. AAV diagnosis. B. AAV treatment. The Y-axis represents the proportion of the topic in the total 3-month corpus.

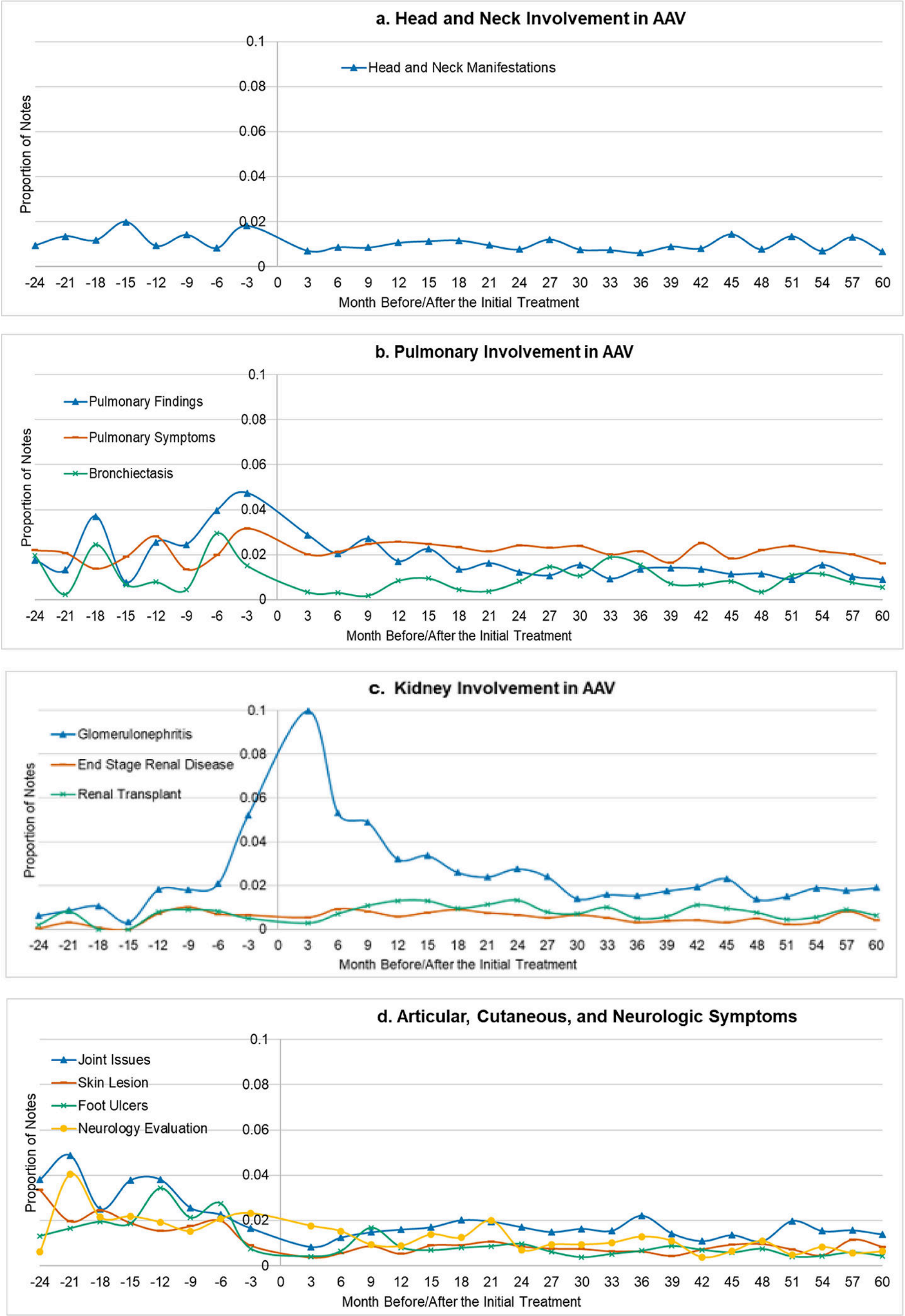


Figure 3.

AAV manifestations. A. Head and neck involvement in AAV. B. Pulmonary involvement in AAV. C. Kidney involvement in AAV. D. Articular, cutaneous, and neurologic symptoms of AAV. The Y-axis represents the proportion of the topic in the total 3-month corpus.

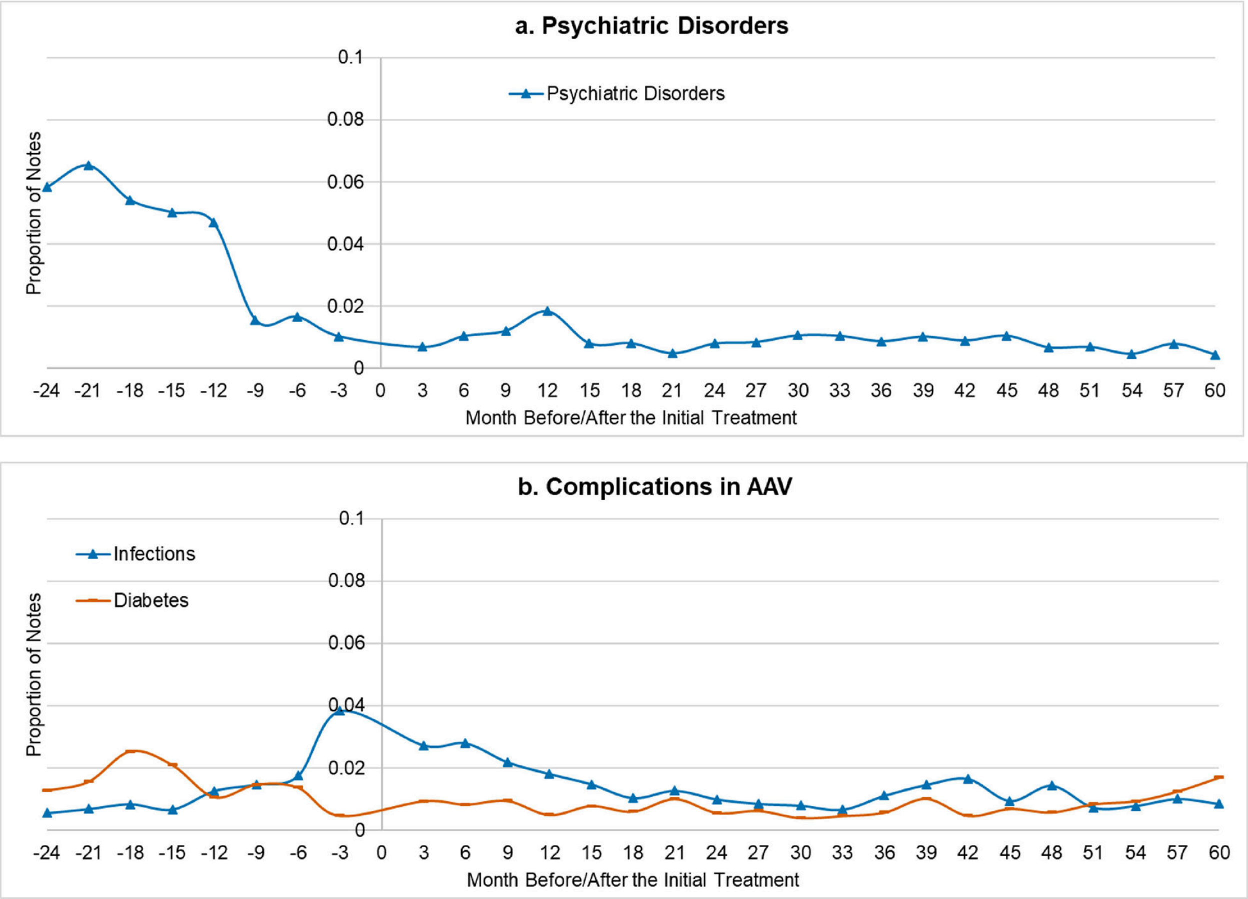


Figure 4.

Complications and comorbidities associated with AAV. A. Psychiatric disorders. B. Complications in AAV. The Y-axis represents the proportion of the topic in the total 3-month corpus.

Table 1.

Characteristics of the Study Cohort

Characteristics	Study Cohort (N=660)
Age at diagnosis, mean (SD), years	56.9 (18.1)
Sex, female, n (%)	392 (59.4)
Race, n (%)	
White	559 (84.7)
Black	18 (2.7)
Other	21 (3.2)
Unknown	62 (9.4)
Ethnicity, n (%)	
Non-Hispanic	598 (90.6)
Hispanic	33 (5.0)
Unknown	29 (4.4)
AAV Disease Characteristics	
Baseline BVAS/WG Score, mean (SD)	4.4 (2.2)
ANCA Type, n (%)	
PR3-ANCA+	268 (41%)
MPO-ANCA+	392 (59%)
Renal Involvement, n (%)	
Any Renal Disease	411 (63%)
End-Stage Renal Disease Ever	105 (16%)
Duration of follow-up, median (IQR), months	118 (61-180)
Prior to the initial treatment	34 (10-69)
After the initial treatment	61 (22-124)
Clinical notes, n (%)	
Prior to the initial treatment	14 637 (12.9)
After the initial treatment	98 411 (87.1)

SD: Standard Deviation; IQR: Interquartile range;

Table 2.

Examples of Stable Topics, Labels, and Analysis Groups.

Groups	Topic Labels	Top 15 Probable Words
AAV Diagnosis	AAV	anca vasculitis rituximab renal skin prednisone ckd rituxan azathioprine wegener chest steroid chronic clear disease
	Granulomatosis with Polyangiitis	anca wegener prednisone azathioprine rituxan clear granulomatosis normal month hypertension stable skin chest past allergy
	Churg-Strauss Syndrome	churg strauss stenosis asthma anca rituximab recurrent wegener cough syndrome steroid dose cell sputum treat
AAV Treatment	AAV-Specific Treatment	tablet anca rituximab prednisone cyclophosphamide steroid vasculitis positive skin chest start normal clear improve month
	General Treatment[*]	prednisone anca rheumatology gpa foot symptom rash swell joint vasculitis deny loss rituximab lab chest
	General Treatment[*]	prednisone swell wegener hand normal joint knee methotrexate esr anca month negative crp vasculitis
Articular, Cutaneous, and Neurologic Symptoms	Skin Lesion	skin lesion area left rash cream dermatology scalp review include face exam apply cell discuss
	Foot Ulcer	grant foot mcg lesion pyoderma premphase dpm refer left toe start gangrenosum ultram normal ditropan
	Joint Issues	left knee hip shoulder joint fracture spine ankle hand foot physical lumbar swell motion mild
	Neurology Evaluation	left normal head intact neurology weakness stroke headache brain nerve seizure exam hand sensation gait
Head and Neck Involvement in AAV	Head and Neck Manifestations	sinus nasal ear left hear nose symptom normal sinusitis strep throat wegener granulomatosis rash pneumo
Pulmonary Involvement in AAV	Pulmonary Findings	chest left lobe lung nodule upper lower find pleural pulmonary small effusion impression opacity pneumonia
	Pulmonary Symptoms	pulmonary lung cough disease chest prednisone anca oxygen dyspnea fev interstitial fvc mild ild increase
	Bronchiectasis[*]	bronchiectasis culture hemoptysis specimen sputum name[†] chest neg neb negative lobe pneumonia pulmonary stable result
	Bronchiectasis[*]	bronchiectasis neg pseudomonas sputum flare oral nasal mucoid chest neb hyperlipidemia inhaler disease sen stable smear
Kidney Involvement in AAV	Glomerulonephritis[*]	renal negative anca urine positive antibody anti igg vasculitis result cell normal protein disease hematuria
	Glomerulonephritis[*]	renal anca vasculitis prednisone cytoxan anemia disease steroid failure creatinine cyclophosphamide admission dose start rpgn
	End Stage Renal Disease	renal esrd fistula transplant left disease folic acid buhle unit catheter nephrocaps access stage hemodialysis
	Renal Transplant	transplant renal tacrolimus prograf donor post cellcept mycophenolate disease bid immunosuppression cmv negative risk rejection
Psychiatric disorders	Psychiatric Disorders	disorder anxiety problem mood depression report support feel axis assessment cocaine current deny think treatment
Complications of AAV	Infections	culture fever negative specimen antibiotic infection sputum infectious result blood admission vancomycin gram report cough

Groups	Topic Labels	Top 15 Probable Words
	Diabetes	unit insulin diabetes give blood lantus sugar dose time nph mellitus scale strip meal slide

[*] Topics with the same labels were merged for trend analysis

[†] a limitation of the name lexicon is that it does not remove all names; Abbreviations: AAV: antineutrophil cytoplasm antibody-associated vasculitides; anca: antineutrophil cytoplasm antibody; mcg: microgram, gpa: granulomatosis with polyangiitis, esr: erythrocyte sedimentation rate, crp: C-reactive protein, neb: nebulizer, fvc: forced vital capacity, fev: forced expiratory volume, mmhg: millimeters mercury, igg: Immunoglobulin-G, rpgn: rapidly progressive glomerulonephritis, esrd: end stage renal disease, cmv: cytomegalovirus, dvt: deep vein thrombosis, ckd: chronic kidney disease