

Towards understanding Neurological manifestations of Lyme disease through a machine learning approach with updated patient registry data

Lorraine Johnson JD MBA, CEO LymeDisease.org;
Lara Kassab PhD, UCLA;
Jingyi Liu, UCLA;
Deanna Needell PhD, UCLA;
Mira Shapiro MSc, Analytic Designers LLC

ABSTRACT

The Centers for Disease Control and Prevention (CDC) estimates that more than 476,000 new cases of Lyme disease occur annually. Many patients remain ill after treatment, and these patients are often referred to as having persistent or chronic Lyme disease (PLD/CLD), and it is estimated that nearly 2 million people in the US have PLD/CLD. PLD/CLD patients suffer from significantly impaired quality of life, utilize healthcare services more frequently and have greater limitations on their ability to work than the general population and patients with other chronic diseases. Neurological symptoms are common in PLD/CLD and have been linked to functional and structural changes in the brain. Very little, however, is known about the neurological manifestations of Lyme disease—contributing factors, diagnostic tools, or best treatments. Here, we analyze the MyLymeData patient registry, a project of LymeDisease.org, that contains over 18,000 patients with Lyme disease. Through statistical analysis and state of the art machine learning (ML) methods, we identify common patterns across patients with neurological symptoms. We compare the statistics across subgroups of neurological symptoms both statistically and through ML topic modeling based on unsupervised non-negative matrix factorization. This research was previously presented at PharmaSUG 2024 and has been updated with a revised sample.

INTRODUCTION

Machine learning (ML) techniques are not a new subdiscipline of Artificial Intelligence (AI) but effectively and credibly using these techniques with real world healthcare data is still in its early stages. Although all the rage in publications and in the media, professionals trained in frequentist and Bayesian statistics are often reluctant to explore these techniques as a valuable tool to improve their understanding of patterns in their patient data.

Our aim in this paper is to provide an example of how ML, specifically unsupervised non-negative matrix factorization (NNMF), may be used in conjunction with standard statistical methods to identify patterns and important variables for use in further analysis. As described in Vendrow et al. (2020), ML may help to elucidate important features that can then be further explored with additional ML experiments and with standard statistical methods. In this example we make use of the patient data contained in the MyLymeData patient registry to explore the manifestations of neurological and other forms of Lyme disease.

DATA & METHODS

We used a sample of over 4,300 patients from the MyLymeData patient registry, a project of LymeDisease.org, that contains over 18,000 patients, to demonstrate how coupling ML with descriptive statistics can result in greater understanding of data patterns and factor relationships.

LIMITATIONS

Although this study includes data from over 4,300 participants in MyLymeData and suggests an exploratory model, some limitations must also be recognized. Participants had access to the Internet and

were not a randomly drawn sample. Those who elect to participate may have been sick longer and more severely ill, which could lead them to seek online support and resources for their illness (Johnson L et al. 2014).

SAMPLE AND INCLUSION/EXCLUSION CRITERIA

We excluded participants who indicated their current Lyme disease status as "well or in remission", "unsure", or "other". Included participants were residents in the United States, diagnosed by a clinician, who described their current stage of Lyme disease as PLD/CLD defined by having remained ill for at least 6 months after treatment with antibiotics for 10 to 21 days. The included participants reported that their disease manifested itself as primarily neurological or primarily musculoskeletal and specified a severity level for at least one of 12 symptoms. As shown in Figure 1, we extracted PLD/CLD and acute patients from the eligible participants as two subgroups for our analysis.

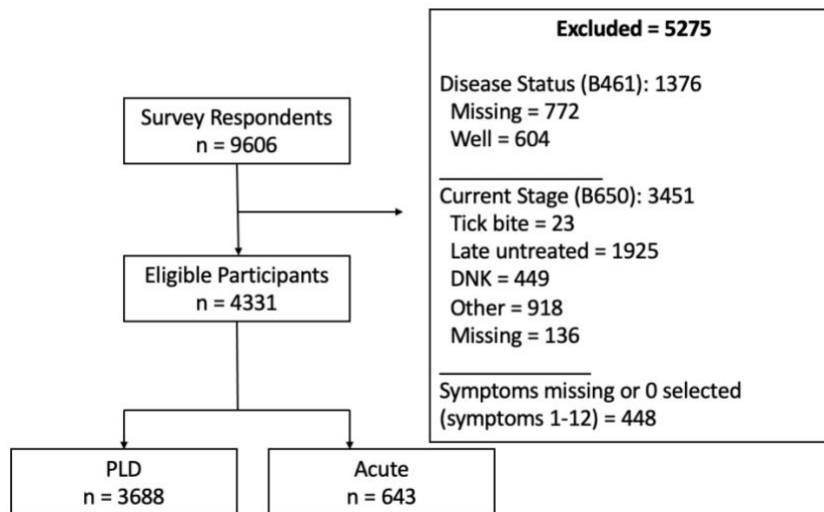


Figure 1 Participant Inclusion/Exclusion Criteria

We asked all participants the question, "How would you describe your current stage of Lyme disease?". We categorized those that answered either "Initial Erythema Migrans (EM) rash (classic bull's eye or irregular rash)" or "Early or disseminated Lyme disease (flu-like, neurologic, cardiovascular or musculoskeletal symptoms)" as acute. Patients who answered "Chronic Lyme disease (remained ill for at least 6 months after treatment with antibiotics for 10-21 days)" were categorized as PLD/CLD. As shown in Figure 1, after meeting the inclusion/exclusion criteria, our sample included 3688 PLD/CLD patients, 1342 who considered their illness primarily neurological and 1440 who considered their illness primarily musculoskeletal (906 PLD/CLD patients who did not answer this question were excluded from our analysis).

Once we isolated our patient sample, we implemented our ML algorithm to identify 4 topics for further exploration. We also computed descriptive statistics on symptom severity for the neurological and musculoskeletal patient subgroups.

ML ALGORITHM DESCRIPTION & IMPLEMENTATION

Non-negative Matrix Factorization (NMF) is a popular ML technique used for topic modeling — the automatic detection of patterns within a data collection [Lee & Seung, 1999]. NMF seeks to identify and extract patterns that are highly exhibited in the data. These patterns typically give us meaningful insights about the population(s) in our data. For example, are there samples that exhibit similar feature

characteristics? If so, what do these feature patterns look like? In the world of topic modeling, a pattern is referred to as a *topic*. NMF is a technique able to automatically identify major feature patterns in the data and associate samples with these patterns.

The input to the NMF algorithm is the data collection in the form of an array (referred to mathematically as a *matrix*), for example, an Excel sheet (or table) with rows representing symptoms, columns representing patients, and entries as the intensity of each symptom given by each patient. NMF takes as input the number of topics desired to detect, the data matrix, and decomposes (or *factorizes*) it into two smaller output matrices. The first matrix, called the *dictionary matrix*, provides the distribution of the symptoms' dominance per topic. The second matrix, called the *representation matrix*, provides the association of the patients to these topics. In our analysis, we normalized the input data matrix.

The name, *Non-negative Matrix Factorization*, refers to the requirement that the two output matrices (the factors) be non-negative. Non-negativity in the decomposition explains the data via additive “building blocks” allowing for ease of interpretation [Lee & Seung, 1999]. NMF has gained in popularity due to large-scale data demands of applications such as document clustering, image processing, financial data mining, audio processing, and genetics.

Many algorithms exist for computing an NMF of a data matrix [Lee & Seung, 2000]. We illustrate how given a data matrix, NMF can be computed using a built-in Python function that is found in the Python sklearn package. Some sample code and the explanation of the parameters we used can be found in Appendix I.

To better illustrate how NMF works, consider our specific study. We used patient responses to questions on the severity of 12 symptoms based on an intensity scale of severe (1- no problem, 2-minimal problem, 3-mild problem, 4-moderate problem, 5-moderately severe problem 6-severe problem, 7-very severe problem). Our goal was to discover whether there were any interesting patterns of symptoms to help understand and define patient groups.

We applied NMF to the data to identify four primary topics (or patterns). These topics are “clusters” of patients that have similar symptom selection profiles. The learned topics numbered from 1 to 4 are visualized as a heatmap below in Figure 2, which displays the dictionary matrix of our factorization. A heatmap is a common tool used to visualize the results of NMF. In this case, the heatmap in Figure 2 serves to visualize what these symptom patterns (presence and intensity) look like. Note in the heatmap, the color scale varies from 0 to 1 (dark to light colors) where lighter colors indicate a higher intensity and greater importance.

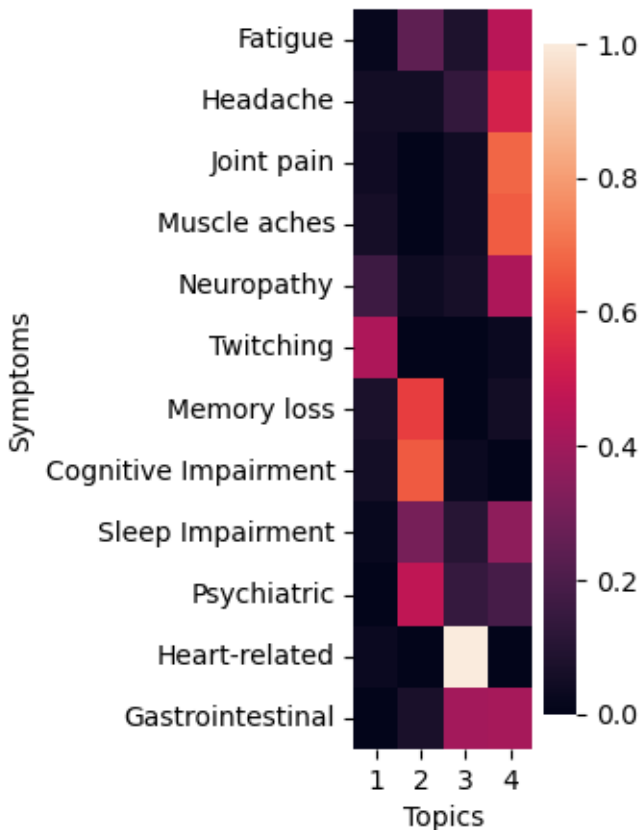


Figure 2. Unsupervised NMF applied to PLD/CLD patients (4 topics)

The unsupervised NMF was used to evaluate which symptoms are grouped under the same topic and which symptoms contributed the most to each topic. As shown in Figure 2, memory loss, cognitive impairment, and psychiatric contributed the most to Topic 2. Twitching was shown to be in its own topic from all the other 11 symptoms. Similarly, heart-related symptom seems to have a pronounced contribution to Topic 3. Multiple symptoms have contributed to Topic 4; however, joint pain and muscle aches are the most important ones.

When observing topics learned by NMF, in general, these are the weights of the features that characterize a topic. So, a higher weight indicates a higher importance of that feature in this topic. In the case where we have a large number of features, we can identify “key” features -- for example, the top 5 features that characterize a topic.

In Figure 2 topics numbered 1-4 are patterns detected by NMF as the most exhibited in the data. The algorithm simultaneously detects how the patients are associated with these topics. For example, a certain patient’s symptoms are highly characterized or explained by the pattern depicted in Topic 2 versus another patient whose symptoms are better represented by a combination of say 80% Topic 2 – 20% Topic 1. So, NMF is also a means of finding how groups within the population possibly cluster in terms of the topics.

DESCRIPTIVE STATISTICS

After implementing the ML techniques, we used traditional statistical methods to compile Table 1. We examined symptom severity in PLD/CLD patients broken down by their self-reported categorization of their disease as primarily musculoskeletal or primarily neurological. For this analysis, to avoid misclassification, we only included patients that answered the question as one of our two categories of interest and excluded those that selected primarily cardiac or other.

Symptom severity is self-reported by patients on a scale from 1 (“no problem”) to 7 (“very severe”). We compared the percentage of patients that reported each symptom as moderate (4) to very severe (7) in each subgroup, musculoskeletal and neurological. In terms of the severity of joint pain, as shown in Table 1, 63.11% of neurological patients described this symptom as moderate or more severe, compared to 84.79% of musculoskeletal patients – a difference of roughly 21%. A similar difference is seen in muscle aches. With respect to neurologic symptoms, more patients in the neurologic subgroup experienced greater memory loss (73.17%), cognitive impairment (75.34%), and psychiatric (70.42%) than musculoskeletal patients.

Almost 87% of neurological and 85% of musculoskeletal patients experienced moderate to very severe fatigue. Memory loss, cognitive impairment, sleep disturbances, and fatigue were rated as moderate to very severe by more than 71% of neurological patients, whereas joint pain and muscle aches were the most severe symptoms for musculoskeletal patients other than fatigue.

	Neurological (1342)		Musculoskeletal (1440)		
Symptoms	% of Total	N	% of Total	N	Difference
Fatigue	86.96%	1167	84.72%	1220	2.24%
Headache	52.46%	704	47.78%	688	4.68%
Joint Pain	63.11%	847	84.79%	1221	-21.68%
Muscle aches	62.15%	834	84.51%	1217	-22.36%
Neuropathy	66.54%	893	62.78%	904	3.76%
Twitching	31.74%	426	25.56%	368	6.18%
Memory Loss	73.17%	982	52.43%	755	20.74%
Cognitive Impairment	75.34%	1011	47.57%	685	27.77%
Sleep Impairment	71.39%	958	64.65%	931	6.74%
Psychiatric	70.42%	945	53.33%	768	17.09%
Heart related	29.36%	394	24.24%	349	5.12%
Gastrointestinal	56.63%	760	53.19%	766	3.44%

Table 1. Percentage and Frequency of Neurological and Musculoskeletal Participants with Moderate to Very Severe Symptoms

RESULTS

As was described, the heatmap generated by applying NMF with 4 topics illustrates the contents of those topics. Of specific interest are these topics: 1 (twitching and neuropathy), 2 (memory loss, cognitive impairment, and psychiatric), and 4 (joint pain and muscle aches). It appears that those symptoms that constitute topic 1 are associated with the peripheral nervous system, topic 2 includes symptoms that usually are considered central nervous system-related, and topic 4 includes symptoms most often

associated with musculoskeletal disease. Several additional investigations and experiments can be performed using NMF. For this set of topics, we might want to explore which patients fall into each topic and examine their characteristics or identify those who self-identified as having an illness that is primarily neurological or musculoskeletal.

However, for this study, we chose to focus on the symptoms associated with the generated topics and explore the symptoms using descriptive statistics and explore the differences between the patient's self-identified illness type, neurological or musculoskeletal. It is interesting that these symptoms of what we might expect to see in an illness that is musculoskeletal seem to correspond with topic 4 in the NMF results. The patients in the neurological group selected moderate to very severe for memory loss, cognitive impairment, and psychiatric with differences from the musculoskeletal group of 20%, 27%, and 17% respectively. These symptoms are what we might expect in patients who have self-identified as having an illness that is primarily neurological in nature and correspond to those symptoms associated with topic 2 in the NMF results. Given that one of our tasks is to define musculoskeletal and neurological subgroups, The confirmation of the ML identified topics of importance with the descriptive statistical approach suggests these groupings are meaningful.

A follow-on experiment using supervised NMF, the neurological / musculoskeletal subgroups, and using the indicator of moderate to very severe for the symptoms could also produce interesting results. In addition, predicting subgroup assignment based on symptom severity could be an additional approach.

CONCLUSION

In this study we have shown how we can use NMF as a first step to identify patterns and associations in a patient sample. These results coupled with descriptive statistics uncovered some possible avenues for understanding this patient population. We have not attempted to prove any associations in this study, but it is clear from these results that patients who experience certain symptoms see their illness as falling into primarily neurological or musculoskeletal. Using a combination of tools and techniques bolstered with researcher subject-area knowledge provides a valuable pathway for moving research forward.

Real world data that is patient-reported, such as in the MyLymeData patient registry, can provide valuable insights to be used in further real-world data studies and to inform future clinical trials. This study demonstrates the potential power that combining ML techniques with traditional statistics can yield, when trying to determine previously unidentified patient subgroups – in this case musculoskeletal vs. neurologic PLD/CLD.

REFERENCES

Johnson, L.; Kassab, L.; Liu, J.; Needell, D.; and Shapiro, M. "Towards understanding Neurological manifestations of Lyme disease through a machine learning approach with patient registry data", PharmaSUG 2024

Lee, Daniel D., and H. Sebastian Seung. "Learning the parts of objects by non-negative matrix factorization." *nature* 401.6755 (1999): 788-791.

Lee, Daniel, and H. Sebastian Seung. "Algorithms for non-negative matrix factorization." *Advances in neural information processing systems* 13 (2000).

<https://scikit-learn.org/stable/modules/generated/sklearn.decomposition.NMF.html#sklearn.decomposition.NMF>

Vendrow, J.; Haddock, J.; Needell, D.; Johnson, L. Feature Selection on Lyme Disease Patient Survey Data. arXiv 2020, arXiv:2019.09087

ACKNOWLEDGMENTS

Thank you to the SESUG Conference Committee for accepting our paper and giving us the opportunity to share our research.

Thank you to the Lyme patients who take their time to share their experience in the MyLymeData patient registry.

For more information please visit: LymeDisease.org

Li, Kassab, and Needell are grateful for the support of NSF DMS 2011140.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Mira Shapiro
mira dot shapiro at analyticdesigners.com