What Do You Think Caused Your ALS? An Analysis of the CDC National Amyotrophic Lateral Sclerosis Patient Registry Qualitative Risk Factor Data Using Artificial Intelligence and Qualitative Methodology

Danielle Boyce, MPH, DPA[1],[2], **Jaime Raymond, MPH**[3], **Theodore C. Larson, MS**[3], **Eddie Kirkland**[4], **D. Kevin Horton, DrPH, MSPH**[3], **Paul Mehta, MD**[3]

[1]Tufts University School of Medicine, Boston, MA USA

[2]Johns Hopkins University Schools of Medicine, Baltimore, MD USA

[3]Centers for Disease Control and Prevention/Agency for Toxic Substances and Disease Registry, Office of Innovation and Analytics, National ALS Registry (CDC/ATSDR), Atlanta Georgia, USA

[4]ICF, Inc. Atlanta, Georgia, USA

Abstract

Objective: Amyotrophic lateral sclerosis (ALS) is an incurable progressive neurodegenerative disease with a significant health burden and poorly understood etiology. This analysis assessed the narrative responses from 3,061 participants in the Centers for Disease Control and Prevention's National ALS Registry who answered the question, 'What do you think caused your ALS?'

Methods: Data analysis used grounded theory qualitative methods and artificial intelligence (AI) using natural language processing (NLP), specifically, Bidirectional Encoder Representations from Transformers (BERT) to explore the participants responses regarding their perceptions for the "cause" of their disease.

Results: Both qualitative and AI analysis methods revealed several, often aligned clusters or themes, which pointed to perceived causes such as genetic, environmental, and military exposures. However, the qualitative analysis revealed detailed themes and subthemes, providing a more comprehensive understanding of perceived causes. Although there were areas of alignment between AI and qualitative analysis, AI's broader categories did not capture the nuances discovered using the more traditional, qualitative approach. The qualitative analysis also revealed that the potential causes of ALS were described within narratives that also sometimes indicated self-blame and other maladaptive coping mechanisms.

Conclusions: This analysis highlights the diverse range of factors that individuals with ALS consider as perceived causes for their disease. Understanding these perceptions can help clinicians

to better support people living with ALS (PLWALS). The analysis highlights the benefits of combining qualitative and AI-based approaches in analyzing narrative data. This rapidly evolving area of data science has the potential to remove barriers to accessing the rich narratives of people with lived experience.

Keywords

Introduction

Amyotrophic lateral sclerosis (ALS) is a fatal, incurable disease that causes progressive degeneration of motor neurons. In the United States in 2018 up to 29,824 persons (9.1 per 100,000 population) were estimated to be living with ALS, with an incidence rate of 1.6 per 100,000 population (1,2). Most people living with ALS (PLWALS) receive a diagnosis 10–16 months after initial symptom onset and survive for 2–5 years after diagnosis (3). ALS and other neurological and psychiatric disorders are attributed to environmental factors and pose a significant health burden to the U.S. population (4–11).

The federal Agency for Toxic Substances and Disease Registry (ATSDR), Centers for Disease Control and Prevention (CDC), established the National ALS Registry (Registry) to better describe the epidemiological trends of ALS in the United States, identify and examine risks and potential causes, and determine the disease's public health burden (12). The Registry collects data from existing national databases and PLWALS who sign up to participate through a voluntary online portal. PLWALS who register gain access to participate in research, receive updates on clinical trials and epidemiological studies, and can donate specimens to the National ALS Biorepository at no cost.

After PLWALS complete the online registry enrollment, they can voluntarily complete up to18 surveys related to demographics and possible risk factors for ALS, from which researchers can request deidentified data to analyze the data for their own research. Participants are also asked two open-ended questions, 'What do you think caused your ALS?' and 'What do you think causes ALS in general?' as one of the surveys. In this analysis, the researchers used traditional qualitative methods and natural language processing (NLP) technology to analyze narrative responses to the first question.

Methods

A narrative structured survey using qualitative questions was one of the survey modules PLWALS can partake in to help researachers to understand their perceptions regarding the "cause" of their disease. Patients who registered in the National ALS Registry, which administers standardized surveys via an online internet portal were included if they completed the qualitative survey. The Registry's methods are previously described (13). Informed consent was obtained under a protocol approved by CDC's Institutional Review Board.

Using the Registry's online portal, participants completed concise surveys on various ALS risk factors and experiences. As of January 2022, participants were offered 18 different survey modules within the web portal (14). These questionnaires were developed and validated by Stanford University's ALS Consortium of Epidemiologic Studies (ACES) (15,16). The surveys are designed to allow respondents to answer the questions independent of a healthcare professional. Participation in the Registry's online portal is entirely voluntary. Some patients sign up for the portal but never fill out any of the surveys, others complete all 18 surveys.

The dataset analyzed in this investigation represents data from 2014 until December 31, 2021 (the date the data were exported from the Registry for analysis). A total of 3,061 Registry respondents completed the survey and were included in the dataset. Participant characteristics including diagnosis year, age at diagnosis, gender, marital status, race, census region, military history, smoking history, alcohol history, family history of ALS, and family history of neurological diseases were also included in the analysis.

Qualitative Analysis

Data were analyzed in an iterative process that drew upon grounded theory (17). Grounded theory uses a 'line by line' coding approach, with each line of the participant narrative captured with initial codes that evolve into focused code. To expedite this approach, a dictionary-based named entity recognition (NER) analysis was performed, coding a '1' for every response that contained certain words or phrases. Although each line was ultimately coded by hand during the qualitative analysis, the NER analysis allowed for quick filtering and sorting of concepts within the same thematic cluster. This was especially useful to quickly classify participant responses consisting only of the phrase 'don't know' and 'no idea,' and to flag responses that might be related to genetics or military experiences. For the AI analysis, all similar themes were placed into clusters.

After initial data analysis and theme development, a modified member checking exercise was performed. Because the principal investigator does not know the identity of the participants in this de-identified dataset, PLWALS and caregivers, outside of the analysis, were informally consulted on themes relevant to their specific areas of interest, including veterans with ALS and people from familial ALS communities. Participants' own responses overwhelmingly agreed with the analysis of responses. Coding of qualitative data and analyses of participant characteristics was performed in SPSS version 27.

Artificial Intelligence Analysis

Artificial intelligence (AI)/NLP technology was also used to perform topic modeling on the open-ended survey responses. The AI analysis was conducted in April 2022 to generate the top 10 topics based on participants responses. Data cleaning, standardization, and topic modelling were performed using open-source libraries in the Python programming language (Python Software Foundation, https://www.python.org/). For data cleaning and standardization, we removed 'stop words', e.g. a, and, the, which are common words that can be safely filtered out without altering the meaning of the text. For stop word removal

we used the Genism stop word library and lemmatized each word via NLTK's WordNet Lemmatizer (18,19).

The cleansed and tokenized text responses were analyzed with Bidirectional Encoder Representations from Transformers (BERT), a machine learning approach for NLP. This was accomplished using the BERTopic library, leveraging the BERT-based *all-MiniLM-L6-v2* sentence-transformers model (20,21). Unigrams, bigrams, and trigrams were all included in the topic modelling process. This analysis resulted in a list of the 10 most prominent topical clusters, based on the vocabulary and syntax of the free-text responses (Table 2).

AI and Qualitative Analysis—Each of the 10 clusters were then labelled based on the predominant themes and keywords used in those clusters (Figure). However, the AI model did not assign a specific label; it simply identified the cluster of responses. The themes were assigned manually based on the keywords and topical themes of the identified cluster. The themes remained broad to eliminate overlapping topics. If similar topics were still in the top 10, they remained as separate topics.

To minimize bias, the qualitative analyst (performed by DB) was blinded to the results of AI analysis (performed by JR and EK) until after the qualitative analysis was complete. A subgroup analysis of the items that were 'ungrouped' or not assigned to one of the top 10 clusters, was then performed, resulting in all 'ungrouped' AI clusters mapping to one or more qualitative themes or subthemes (Table 3).

Results

Demographic Characteristics

The analysis included 3,061 people living with ALS. Participant characteristics recorded on close-ended question responses included diagnosis year, age at diagnosis, gender, marital status, race, census region, military history, smoking history, alcohol history, family history of ALS, and family history of neurological diseases (Table 1). The distribution of diagnosis years ranged from before 2011 to 2021. Most participants were diagnosed between 2013 and 2019, with the highest proportion occurring in 2014 (457 participants; 14.9%). Most of the participants were male (1,790; 58.5%), married (2,418; 79.2%), and White (2,940; 96.0%). The most common census region was Region 3, the southern United States (1,118 participants, 37.0%). Most participants were aged 60–69 years (1,175; 38.4%) or 50–59 years (873; 28.5%) at diagnosis.

The Registry also collects data on participants' occupational and lifestyle characteristics. In this analysis, 583 participants (19.1%) had a military history, 1,311 participants (42.8%) had ever smoked, and 2,467 participants (80.6%) had ever consumed alcohol (Table 1). Most participants did not have a family history of ALS (2,862; 93.5%) or a family history of neurological diseases (2,484; 81.5%).

AI Generated Cluster Analysis

When considering clustering of responses, using AI classification described above, most participant responses did not fit into a specific AI cluster (60%) and were classified as

'ungrouped' (Table 2). Among the identifiable topics, the most common response was classified as 'unknown topic' (16.3%). The next most common topic of responses were related to family history of other neurological diseases (4.5%) and heredity (4.4%) (Table 2).

AI and Manual Thematic Analysis of Perceived Cause of ALS

Table 3 compares the major themes emerging from the AI and traditional qualitative analysis, regarding perceived cause of ALS, with a few representative participant quotations. The major difference was that the manual thematic analysis grouped responses into themes and subthemes, which the AI/NLP analysis did not. This allowed for all the 'ungrouped' responses identified by AI/NLP to be labeled in the manual thematic analysis process.

The AI cluster of responses around perceived "cause" aligned with the qualitative themes and subthemes in some areas, and not in others (Table 3). For example, AI identified a cluster of responses called 'Chemical exposure,' which corresponds to the 'Environment/ Exposure to chemicals/Pesticides' theme in the qualitative analysis. The AI model identified a 'Diet/Exercise' cluster as a potential perceived "cause" of ALS, which corresponds to subthemes within the 'Lifestyle' theme in the qualitative analysis. The subthemes of drugs/ alcohol and exercise/sports/athletics/heavy physical labor in the qualitative analysis are also aligned with the AI clusters. The AI approach identified a cluster of responses as 'Family history of other neurological disease,' which had a nearly identical corresponding theme in the qualitative analysis.

The AI model also identified a cluster of 'Head Trauma' responses as a potential perceived "cause" of ALS, whereas the corresponding qualitative analysis theme was 'Accident/Injury' with 'Head injury' as a subtheme. The AI cluster and qualitative analysis both identified 'Genetic(s)' as a potential perceived "cause" of ALS; the qualitative analysis included subthemes to differentiate general versus specific genetic etiological concepts, which were not specifically identified by the AI clustering method. AI modelling and qualitative analysis both identified 'Military' as a perceived "cause" of ALS. The subthemes of medications and immunizations, occupational exposures (including burn pit and nuclear radiation exposure), and Agent Orange exposure in the qualitative analysis are more specific than the AI clustering method's broader category of 'Military.'

Discussion

In this analysis, we explored the responses of people with ALS about their perspectives on the potential "causes" of their condition. By analyzing the free text responses of people living with ALS, we sought to gain a more comprehensive understanding of the life experiences of those living with ALS and uncover potential factors that might have contributed to the development of the disease. The responses provide valuable insights into the participants' beliefs about the "cause" of their ALS and highlight the importance of considering participants' perspectives in ALS research.

One of the key findings from the patient responses was the diversity of opinions regarding the perceived potential "causes" of ALS. This variation in beliefs underscores the complex and heterogeneous nature of ALS. The narrative responses analyzed in this analysis provide

unprecedented detail on specific incidents regarding previously identified exposures, such as military service, environmental exposures, and occupational exposures. These details might provide important, hypothesis-generating data that could not be gleaned from a multiple-choice survey question.

The responses from PLWALS revealed the emotional and psychological impact of living with a poorly understood and incurable condition. Many patients expressed sentiments related to frustration, confusion, helplessness, and maladaptive coping, including self-blame. Self-blame is most frequently found in the 'Lifestyle factors' theme, such as diet, exercise, illicit drugs, tobacco use, and alcohol. Maladaptive coping is known to be associated with poorer patient-reported outcomes in other diseases (22). The concerning sentiments shared by Registry respondents might contribute to higher levels of depression and anxiety, which are related to poor health-related quality of life among people with ALS (23).

The themes developed and sentiments discovered in this analysis present opportunities to translate results into improved support for PLWALS and their families. Clinicians and therapists can use this information to create tailored and empathetic approaches to working with people living with ALS, which can help reduce feelings of self-blame and empower their patients to better cope with the challenges of the disease. Understanding the importance of these lifestyle factors in patients' minds can enable clinicians and therapists to develop supportive interventions that promote better coping strategies, providing evidence-based information that helps to alleviate feelings of guilt and reinforce the complex and multifactorial nature of the disease. By acknowledging and addressing patients' beliefs and fears about the causes of their ALS, healthcare professionals can strengthen the patient-provider relationship and contribute to improved mental well-being and overall quality of life for those living with this devastating disease.

Limitations

This analysis has several limitations. It is a voluntary survey. Not everyone who registers completes the surveys including the open-ended survey on which this analysis is based. Participants with internet access are presumably more likely to participate; this might skew the population toward a younger, better educated patient sample. The portion of younger participants (ages 40–49 years) is overrepresented in this sample (0.6%) as compared to national prevalence in the Registry of 8.3% (23). The oldest age group, 80 years and older, is underrepresented in this analysis (2.4%) compared with data in the National ALS Registry (8%). At 4.0%, non-White race is underrepresented in this sample compared with 11.9% in the Registry as a whole (23). Potential reasons for these discrepancies include barriers to accessing the technology needed for self-registration as we only have patients in the Registry that have this access; lack of awareness of the Registry which could be due to lower utilization of ALS specialty clinics; and reduced participation by residents of the Western United States, a region with a substantial non-White population (2).

Many free-text responses were labelled as 'ungrouped' in the topic in the AI methodology modelling process, indicating that the model was not able to assign those data points exclusively to any of the identified topics. This could happen for several reasons. The generic sentence-transformers model was not fine-tuned, and as such, it might have had

difficulty generalizing to this specific use case and assigning the ALS-related responses to a specific label during the AI methodology modeling process. This could be due to irrelevant language in the free-text responses that confuses the generically trained model. Novel tokens or meaningless ('noisy') data might result in specific posts being labelled as 'ungrouped' despite the presence of otherwise categorizable information. For example, the BERT model might have difficulty assigning the comment 'Had shingles twice. Also spent 20+ years around cat litter.' to a specific topic. This analytic challenge is due to several factors, including ambiguity, limited context, or ill-formed sentence structure. Ambiguity plays a role in ungrouping because the comment contains two distinct statements that might not be directly related to each other. The BERT model will probably have difficulty determining a common context for the two statements, which are disparate and seemingly unrelated to ALS. As a result, the model will struggle to assign this post to a specific topic cluster. The limited context of this response also might pose a challenge. The comment is very concise and contains no mention of a relationship to ALS. Transformers-based models such as BERT rely on language context and surrounding information to generate embeddings for a given group of tokens. Because of that, this post might not contain enough textual context to accurately label the response within a specific topic.

Tokenization and sentence structure also might play a role in topic ambiguity. Using a language model requires breaking statements into tokens (sub-words, words, or multi-word chunks) and analyzing their relationships. A comment's structure, with two separate statements syntactically joined by 'also,' might affect the model's ability to understand the intended meaning. If the tokenization and sentence structure do not align well with the patterns that the BERT-based model has learned, it might not group the comment correctly.

Topic modelling performance might be improved by fine-tuning the underlying model on a specific dataset or task, such as a curated dataset of comments related to ALS. This might help BERT, or another transformers-based model, to better understand the context and associations in the data and improve its ability to group comments such as the one mentioned. Topic modelling also might be improved by using the embeddings of large language models (LLMs), which have been released since this study and have seen significant gains in generalizing to unseen text. Although the fine-tuning process was outside the scope of this analysis, and many of the high-performing LLMs were not available at the time of this analysis, they represent promising avenues for future research. For this analysis, given the large number of ungrouped responses, it was important to assess those data points separately to determine whether they represent important information that was missed by the model or by the data pre-processing steps.

Conclusion

This is an analysis of a large qualitative dataset with AI and traditional qualitative approaches. Although the qualitative approach resulted in a more comprehensive theme and subtheme development, AI provides an unbiased, reinforcing check for the traditional qualitative analysis. AI also zooms in on specific topics that might be important but not given the same prioritization in traditional thematic analysis. By combining AI and

traditional qualitative analytical techniques, researchers are sure to leave no stone unturned in the quest for the most accurate and actionable characterization of the data.

The narrative responses housed in the National ALS Registry represent the largest collection of lived experience data for people with ALS in the world and provide insight into the participants' theories about the "causes" of their ALS. Engaging with people with ALS and asking for their perspectives on the potential "causes" of their disease offers numerous benefits. Those include gaining a deeper understanding of the complex factors contributing to ALS onset and progression, fostering a more empathetic and supportive clinical and research environment, and leveraging patient insights to drive research and generate hypotheses. It is important for scientists to tap into the opportunities presented by this rich dataset. This analysis also demonstrates that, with recent technological advances, it is possible to use the same tools that we used in this analysis to analyze large narrative datasets on laptops with open-source technology. This analysis highlights the value of having a partnership between AI and human analysis. Insights gleaned from our experimentation with different approaches to the analysis of these large, unstructured data will be helpful to scientists in all disease areas. Future studies should continue to prioritize patient engagement and incorporate their perspectives into the research process to advance our understanding of ALS and improve patient care.

Acknowledgements:

Data Availability Statement:

The data are available via CDC - Amyotrophic Lateral Sclerosis: Research Application Form.

References

1. Mehta P, Raymond J, Punjani R, Larson T, Han M, Bove F, Horton DK. Incidence of amyotrophic lateral sclerosis in the United States, 2014–2016. Amyotroph Lateral Scler Frontotemporal Degener. 2022;23:1–5.

2. Mehta Paul, Raymond Jaime, Zhang Yuzi, Punjani Reshma, Han Moon, Larson Theodore, Muravov Oleg, Lyles Robert H. & Horton D. Kevin (2023) Prevalence of amyotrophic lateral sclerosis in the United States, 2018, Amyotrophic Lateral Sclerosis and Frontotemporal Degeneration, DOI: 10.1080/21678421.2023.2245858

3. Richards D, Morren JA, Pioro EP. Time to diagnosis and factors affecting diagnostic delay in amyotrophic lateral sclerosis. J Neurol Sci. 2020;417:117054.

4. Andrew AS, Bradley WG, Peipert D, Butt T, Amoako K, Pioro EP, et al. Risk factors for amyotrophic lateral sclerosis: a regional United States case-control study. Muscle Nerve. 2021;63(1):52–59.

5. Goutman SA, Boss J, Patterson A, Mukherjee B, Batterman S, Feldman EL. High plasma concentrations of organic pollutants negatively impact survival in amyotrophic lateral sclerosis. J Neurol Neurosurg Psychiatry. 2019;90(8):907–912.

6. Cox PA, Kostrzewa RM, Guillemin GJ. BMAA and neurodegenerative illness. Neurotox Res. 2018;33(1):178–183.

7. Bozzoni V, Pansarasa O, Diamanti L, Nosari G, Cereda C, Ceroni M. Amyotrophic lateral sclerosis and environmental factors. Funct Neurol. 2016;31(1):7–19.

8. Weisskopf MG, Cudkowicz ME, Johnson N. Military service and amyotrophic lateral sclerosis in a population-based cohort. Epidemiology. 2015;26(6):831–838.

9. Capozzella A, Sacco C, Chighine A, Loreti B, Scala B, Casale T, et al. Work related etiology of amyotrophic lateral sclerosis (ALS): a meta-analysis. Ann Ig. 2014;26(5):456–472.

10. Yu Y, Su FC, Callaghan BC, Goutman SA, Batterman SA, Feldman EL. Environmental risk factors and amyotrophic lateral sclerosis (ALS): A case-control study of ALS in Michigan. PLoS One. 2014;9(6):e101186.

11. Cox PA, Richer R, Metcalf JS, Banack SA, Codd GA, Bradley WG. Cyanobacteria and BMAA exposure from desert dust: a possible link to sporadic ALS among Gulf War veterans. Amyotroph Lateral Scler. 2009;10 Suppl 2:109–117.

12. US Public Health Service. ALS Registry Act. Washington, DC: 110th Congress; 2008.

13. Antao VC, Horton DK., The National Amyotrophic Lateral Sclerosis (ALS) Registry. J Environ Health. 2012;75(1):28–30.

14. Cedarbaum JM, Stambler N, Malta E, Fuller C, Hilt D, Thurmond B, Nakanishi A, BDNF ALS Study Group (Phase III). The ALSFRS-R: a revised ALS functional rating scale that incorporates assessments of respiratory function. BDNF ALS Study Group (Phase III). J Neurol Sci. 1999;169(1–2):13–21.

15. Horton DK, Mehta P, Antao VC. Quantifying a nonnotifiable disease in the United States: the National Amyotrophic Lateral Sclerosis Registry model. JAMA. 2014;312(11):1097–1098.

16. US Department of Health and Human Services. HHS regional offices. Washington, DC: US Department of Health and Human Services; 2017. https://www.hhs.gov/about/agencies/iea/regional-offices/index.html

17. Charmaz K. Constructing grounded theory: a practical guide through qualitative analysis. Thousand Oaks (CA): Sage Publications; 2006.

18. Gensim. parsing.preprocessing—Functions to preprocess raw text. 2022. [accessed 2023 May 26]. https://radimrehurek.com/gensim/parsing/preprocessing.html.

19. Project NLKT. Documentation: Source code for nltk.stem.wordnet. 2023. [accessed 2023 May 27]. https://www.nltk.org/_modules/nltk/stem/wordnet.html.

20. Grootendorst M. BERTopic. Neural topic modeling with a class-based TF-IDF procedure. arXiv preprint. arXiv:2203.05794. 2022. https://maartengr.github.io/BERTopic/index.html#citation.

21. Riemers N. Pretrained models. SBERT.net. 2022. https://www.sbert.net/docs/pretrained_models.html.

22. Chao CY, Lemieux C, Restellini S, Afif W, Bitton A, Lakatos PL, Wild G, Bessissow T. Maladaptive coping, low self-efficacy and disease activity are associated with poorer patient-reported outcomes in inflammatory bowel disease. Saudi J Gastroenterol. 2019;25(3):159–166.

23. van Groenestijn AC, Kruitwagen-van Reenen ET, Visser-Meily JM, van den Berg LH, Schröder CD. Associations between psychological factors and HRQoL and global quality of life in patients with ALS: a systematic review. Health Qual Life Outcomes. 2016;14(1):107. doi: 10.1186/s12955-016-0507-6.
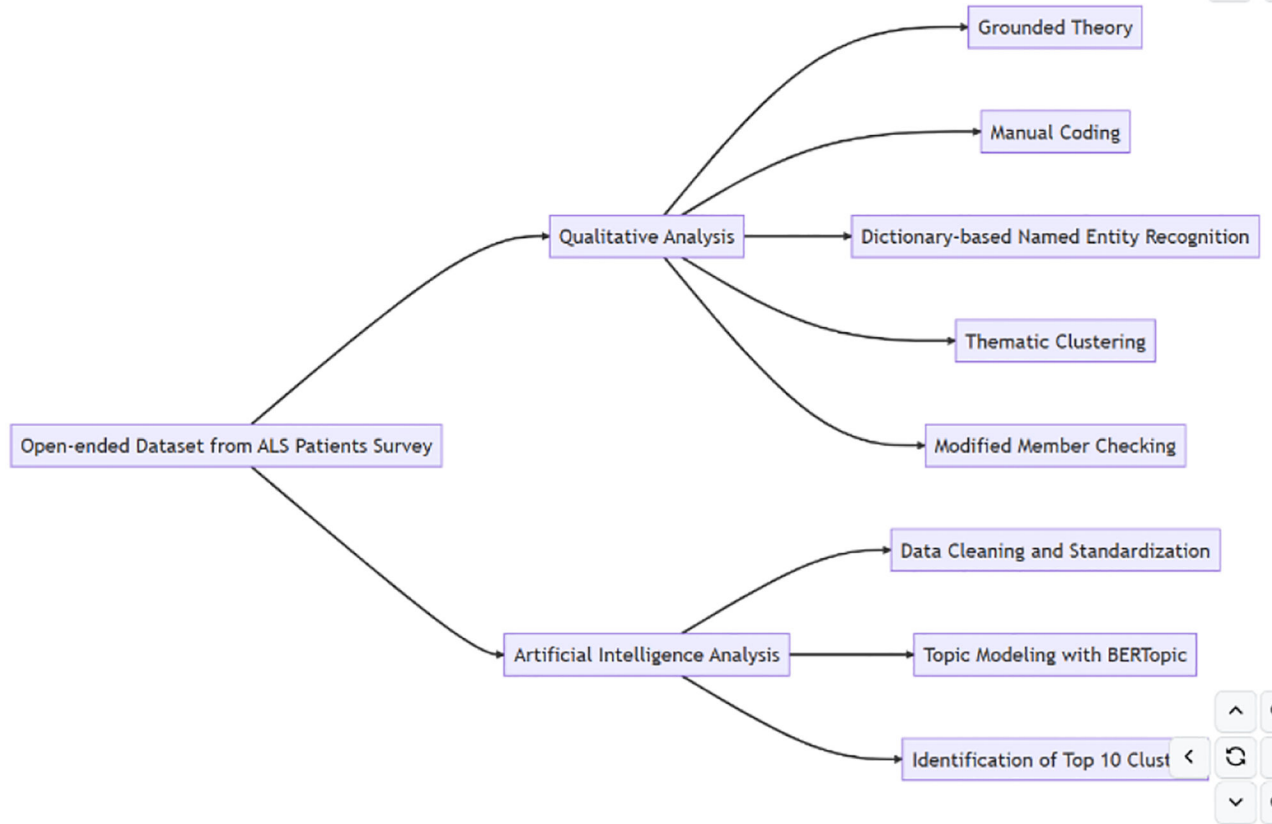
Figure.
Creating themes using the survey reponses.

Table 1.

National Amyotrophic Lateral Sclerosis (ALS) Registry participant characteristics (N = 3,061)

Diagnosis year	n	%
Pre-2011	210	6.9
2011	89	2.9
2012	166	5.4
2013	382	12.5
2014	457	14.9
2015	348	11.4
2016	335	10.9
2017	261	8.5
2018	282	9.2
2019	263	8.6
2020	176	5.8
2021	92	3.0
Diagnosis age (yrs)		
<40	93	3.0
40–49	323	10.6
50–59	873	28.5
60–69	1,175	38.4
70–79	525	17.2
80	72	2.4
Gender		
Male	1,790	58.5
Female	1,271	41.5
Marital status		
Never married	134	4.4
Married	2,418	79.2
Separated	26	0.9
Divorced	293	9.6
Widowed	121	4.0
Living with partner	63	2.1
Race		
White	2,940	96.0
Non-White	121	4.0
Census region*		
1	403	13.4
2	833	27.6
3	1,118	37.0
4	665	22.0

Diagnosis year	n	%
Military history		
Yes	583	19.1
No	2,478	80.9
Smoking history		
Never smoked	1,750	57.2
Ever smoked	1,311	42.8
Alcohol history		
Never drank	594	19.4
Ever drank	2,467	80.6
Family history of ALS		
No	2,862	93.5
Yes	199	6.5
Family history of neurological disease		
No	2,484	81.5
Yes	565	18.5

*Northeast: Connecticut, Maine, Massachusetts, New Hampshire, New Jersey, New York, Pennsylvania, Rhode Island, Vermont; South: Alabama, Arkansas, Delaware, District of Columbia, Florida, Georgia, Kentucky, Louisiana, Maryland, Mississippi, North Carolina, Oklahoma, South Carolina, Tennessee, Texas, Virginia, West Virginia; Midwest: Iowa, Illinois, Indiana, Kansas, Michigan, Minnesota, Missouri, Nebraska, North Dakota, Ohio, South Dakota, Wisconsin; West: Alaska, Arizona, California, Colorado, Hawaii, Idaho, Montana, Nevada, New Mexico, Oregon, Utah, Washington, Wyoming.

Table 2.

Topics identified by artificial intelligence for responses to question, 'What do you think caused your ALS?'

Topic Clusters from responses	n	%
Ungrouped	1,859	60.7
Unknown	499	16.3
Family history of other neurological disease	137	4.5
Heredity	135	4.4
Diet/Exercise	69	2.3
Head trauma	57	1.9
Genetics	57	1.9
Military	59	1.9
Concussion	48	1.6
Statins	48	1.6
Stress	47	1.5
Chemical exposure	46	1.5

*
Ungrouped cluster reflects responses that did not fit into the top responses.

Table 3.

Comparison of top artificial intelligence (AI) topics and qualitative analysis themes, with sample responses from persons with ALS.

AI topic cluster	Corresponding qualitative analysis main theme	Qualitative analysis subtheme 1	Qualitative analysis subtheme 2	Sample response
Chemical exposure	Environment/exposure to chemicals/pesticides		Broadly defined, no specific incident	"modified food, air quality, water quality…'
	Environment/exposure to chemicals/pesticides		Specific incident recalled	"Possible exposure to antifungal or pesticides in yard o exposure to chlorine (swimming pools)."
	Exposure to metals/mercury			"I played with mercury from broken thermometers occasionally as a child."
	Occupational/non-agricultural, non-military		Broad/Long-term exposures	"The exposure to non-lead and lead paint happened while marking timber working for the US Forest Service. The lead based paint was used I believe until the early to mid 1990s. I started in 1987 so I was exposed to lead based paint for a few years. We would not use any facemask or protective breathing device and many times our faces would be the color of the paint we would be using, inhaling the paint."
			Firefighting	"There are several factors I believe could have caused The ALS. The first would be that I was a firefighter for 30 years I was exposed During numerous fires to toxic smoke from structure fires vehicle fires and wildland fires. During the early part of my career we did not have and use breathing apparatus like we do in modern firefighting. Even today after 30 years of firefighting there is still no breathing apparatus for wildland firefighters."
	Agricultural/farming			"For me, it was doing a lot of manual labor work on my new farm, fencing, using chemicals, and working too hard every evening and after work."
Diet/Exercise	Lifestyle	Drugs/alcohol		"Perhaps my beer drinking (2–3 beers/day), but I don't like to get drunk. For now, the past few months, I have given up beer. I don't drink wine or hard liquor."
		Smoking		"I was a moderate smoker (5–10 cigarettes per day) from age 19 to age 39."
		Diet		"I also drank 5–6 diet sodas every day for 25 years. I have to wonder if those chemicals did something."
		Exercise/sports/athletics/heavy physical labor		"12 hour of exercise in a gym per week"
Family history of other neurological disease	Family history of other neurodegenerative disease			"I am wondering if my Mother's Dementia (Alzheimers?) and my brother's Parkinson Disease may be related to my ALS diagnosis. 3 of Mother's sisters also had Dementia."
Head trauma	Accident/Injury	With reference to head injury		"…I went over the handlebars on my mountain bike and landed squarely on my head. I hurt my neck quite a bit, but did not seek medical attention. I really thought that I injured something."
		With no reference to head injury		"I have had several traumatic accidents in my life, I had an accident at work where a stack of 700–800 lbs of cardboard boxes fell on my right leg breaking it in 3 places. I was hit by a train in 2004. I walked away with that with only a sprained ankle."
Concussion				*See head trauma*

AI topic cluster	Corresponding qualitative analysis main theme	Qualitative analysis subtheme 1	Qualitative analysis subtheme 2	Sample response
Genetics	Genetic	General theory of genetic basis for disease with no known family history		
		Known genetic cause/family history-testing or gene name was referenced		"I have a mutated SOD1 gene. Three other members of my family have had ALS."
		Genetic cause - relatives w/ALS or ALS-like disease but no evidence of genetic testing		"caused by familial; possibly from the chemicals I worked with and also the various armaments that the military has (including nuclear)."
Heredity				*See Genetics*
Unknown	Don't know/none			"I don't know." "Who knows." "I wish I knew."
Military	Military	General reference to military service-related exposures		"I can assume it was caused when I deployed to desert storm. I was exposed to burning oil wells, burn pits, drove by vehicles that were destroyed with munitions, setup outside were the Army destroyed a cache of ordnance and took some pills directed by my superiors for possible chemical attack."
		Immunizations, medications received in military		"I can only speculate that heavy exposure to agent orange in Vietnam, or the rapid vaccination process prior to going to Vietnam, where vaccines were given for a variety of diseases within a 2-week period and should have been spread over a 6 months period, may have been two trigger points for ALS later in life. The vaccines given were for standard immunizations as well as for overseas immunizations for typhoid, paratyphoid, yellow fever, cholera, and plague. While in #ry we were given gamma-globulin and chloroquine. The whole combination is a witches' brew."
		Agent Orange		"While in the USAF I was exposed to a variety of different chemicals, fungicides, fuels and classified item, in a confined cargo bay for numerous hours in flight. Also exposed to Agent Orange and other chemicals while serving time in Vietnam."
		Occupational exposure	Burn pit	"I think it had to do with deployment to Iraq. I never had blood pressure problems until that deployment. Was in #ry 14 months. We had burn pit on site and i ran the fuel truck with jp8 everyday."
			Nuclear/ radiation	"I was stationed at a fort guarding New York Harbor that had a Nike Missile site, with Hercules missiles. It was reputed that they rotated nuclear missiles through at times. My job required me to go all over the post as I was involved in keeping the post running for the Air Defense Artillery unit. It was a very old fort that had old battery installation too. We skated on the cooling ponds."
Statins	Personal medical history	Medications, non-statins		"I really think what caused it was too much blood pressure medication. Many tests such as CT scan, MRI, for my heart. Also being given Ciprofloxacin. Also I had a nuclear stress test. After all of these things I started having the symptoms of ALS."
		Statins		"I took High dose statins for more than 10 years…I think the statin use caused myopathy and neuropathy leading to ALS."
		Surgery or other medical procedure		"I had a total knee surgery in September. I started showing signs in November/December."
		Chronic comorbidities such as diabetes, arthritis		"The year I had first symptom of ALS I had shingles earlier that year and took medication for the shingles. Also have type two diabetes, and was given insulin that caused extremely low blood sugar. I wonder if those caused it."

AI topic cluster	Corresponding qualitative analysis main theme	Qualitative analysis subtheme 1	Qualitative analysis subtheme 2	Sample response
		Virus, like EBV, STDs, meningitis, excluding Lyme		"I have familial ALS- SOD-1 G94A. As for triggers, I had mononucleosis several years prior to developing ALS. My mono caused viral meningitis and I have had low to moderate levels of EBV (Epstein Barr Virus) since recovering from mono/meningitis."
		Lyme disease		"I have also been around pets and ticks. I wonder how Lyme Disease might cause, mimic or contribute to ALS."
		Immunizations–not military-related		"I had a vaccine adverse reaction injury caused by my flu vaccine…I believe this triggered my ALS."
Stress	Emotional stress or trauma			"I believe an extreme amount of stress caused this. I was working with my business partner for several years with great difficulty. Then one day I had had enough and told him I am retiring. This was in the beginning of January and by the end of the month I started with slight drop foot. I had also lost my daughter two years prior to that."
Ungrouped	Other			"Na" "I have ALS4." "It is part of God's plan for my life."
No corresponding AI theme identified	Nuclear/radiation exposures/non-military			"grew up fifty mile from Nevada above ground nuclear testing site"

*
Manual thematic analysis grouped responses into themes and subthemes, which the AI/NLP analysis did not. This allowed for all the 'ungrouped' responses identified by AI/NLP to be categorized.