www.nature.com/scientificreports

scientific reports

OPEN

Natural language processing reveals network structure of pain communication in social media using discrete mathematical analysis

Nobuo Okui[1,2,3] & Shigeo Horie[1,4]

Pain-related discussions on social media provide valuable insights into how people naturally express and communicate their pain experiences. However, the network structure of these discussions remains poorly understood. This study analyzed 57,000 Reddit comments from the GoEmotions dataset (2005–2019) using natural language processing and network analysis techniques grounded in discrete mathematical principles. The constructed network, comprising 5,630 nodes and 86,972 edges, revealed complex patterns of pain-related language use. The network exhibited a sparse overall density (0.0055) but a high clustering coefficient (0.7700), indicating the presence of distinct thematic communities. At the center of the network was the term *pain*, which showed the highest degree centrality (0.821429), reflecting its semantic anchoring function in pain discourse. Other terms, such as *headache*, served as context-sensitive bridge nodes that connected different semantic subdomains. In contrast, terms like *burning*, despite moderate centrality values, were found to co-occur predominantly with metaphorical or decorative expressions rather than emotion- or symptom-related descriptors. Community detection revealed 12 distinct clusters, with the largest containing 1,021 nodes, capturing diverse aspects of pain communication. Stability analysis demonstrated that core pain-related terms maintained consistent centrality, while peripheral or metaphorical terms showed greater variability. These findings offer novel insights into the semantic structure of pain discourse and suggest that network analysis of social media discussions can inform improved clinical communication and symptom assessment.

Keywords Pain perception, Digital health communication, Discrete mathematics in health communication, Natural language, Pain assessment in social media, Symptom network analysis

Various methods have been used to study emotion. Research based primarily on data obtained from surveys and controlled experimental environments has been common, employing techniques where participants evaluate stimuli, such as photographs of facial expressions or music, to explore emotional dimensions and experiences[1,2]. However, these methods have limitations in terms of ecological validity, that is, the extent to which the findings reflect natural language use and emotional expression in real-life contexts. Studies based on natural text expressions found on social media or online forums, which are part of daily communication, remain limited[3,4].

Unlike other emotions, *pain* is not only a biological or psychological phenomenon but also a communicative act deeply embedded in social and intersubjective contexts. The Social Communication Model of *Pain* posits that linguistic expression plays a critical role in shaping how *pain* is perceived, evaluated, and responded to by others[5]. Phenomenological approaches further emphasize that *pain* is lived and expressed within a "lifeworld," where metaphor and imagination help constitute its meaning beyond numerical intensity[6,7]. These perspectives suggest that the study of *pain*-related language can illuminate the social construction of suffering and the dynamics of its recognition.

[1]Data Science and Informatics for Genetic Disorders, Juntendo University, 2-1-1 Hongo, Bunkyo-ku, Tokyo 113-8421, Japan. [2]Data Science, Kanagawa Dental University, Inaoka-cyou 82, Yokosuka, Kanagawa 238-0008, Japan. [3]Urology, Yokosuka Urogynecology and Urology Clinic, Yokosuka 238-0008, Japan. [4]Urology, Graduate School of Medicine, Juntendo University, Tokyo 113-8421, Japan. email: okuinobuo@gmail.com

This study analyzed the network structure of discussions related to *pain* on online platforms using Natural Language Processing (NLP) techniques. Unlike other emotions, *pain* is a complex experience that encompasses both psychological and physiological effects, positioning it uniquely in emotional research[8,9]. The GoEmotions dataset enabled the construction of a *pain*-related terms network, with topological analysis revealing how *pain* discussions unfolded[10]. This approach offers a novel perspective for understanding the experience of *pain* through natural language expressions, independent of self-reports or survey-based methods[11].

Results

Network structure and central nodes

This section provides a global overview of the *pain*-related lexical network, focusing on its structural scale, density, and community structure derived from 123,840 word co-occurrence relations. The resulting network comprises 5630 nodes and 86,972 edges, where each node represents a word and each edge denotes co-occurrence[4,8,9]. Terms such as *burning*, *headache*, *discomfort*, and *ache* exhibit prominent connectivity, suggesting frequent semantic associations.

Figure 1 illustrates a core–periphery structure: larger central nodes reflect higher centrality and discourse influence, while peripheral nodes reveal the stratified nature of *pain*-related language[11,12].

Although sparse in overall density (0.005500), the network displays high local connectivity. The average degree is 30.900000, indicating that each term connects to roughly 31 others[13]. The network diameter is 5, suggesting even distant terms are linked via short paths, supporting efficient semantic flow[14].

The clustering coefficient of 0.770000 confirms that words form tightly connected local subgroups[15]. Louvain community detection identified 12 distinct communities[16], with the largest including 1,021 nodes and others containing 911, 842, 520, and 495. These findings indicate a globally cohesive structure with semantically distinct subgroups reflecting thematic and contextual variation.

Structural roles of pain-related terms

This section focuses on identifying the lexical roles and relative centrality of individual symptom-related terms within the network. Figure 2 shows centrality metrics for key pain-related terms and their co-occurrence patterns. The analysis revealed a structured subgraph of 309 nodes and 363 edges[11,12], capturing the complexity of symptom language on social media, including both frequent and unique word pairings.

The term *pain* consistently scored highest across all three centrality metrics—degree (0.821429), betweenness (0.930134), and eigenvector centrality (0.695893)—indicating its dominant role in organizing and connecting discourse elements across the lexical network[13,15,17]. In contrast, terms such as *headache* (0.107000, 0.109000,

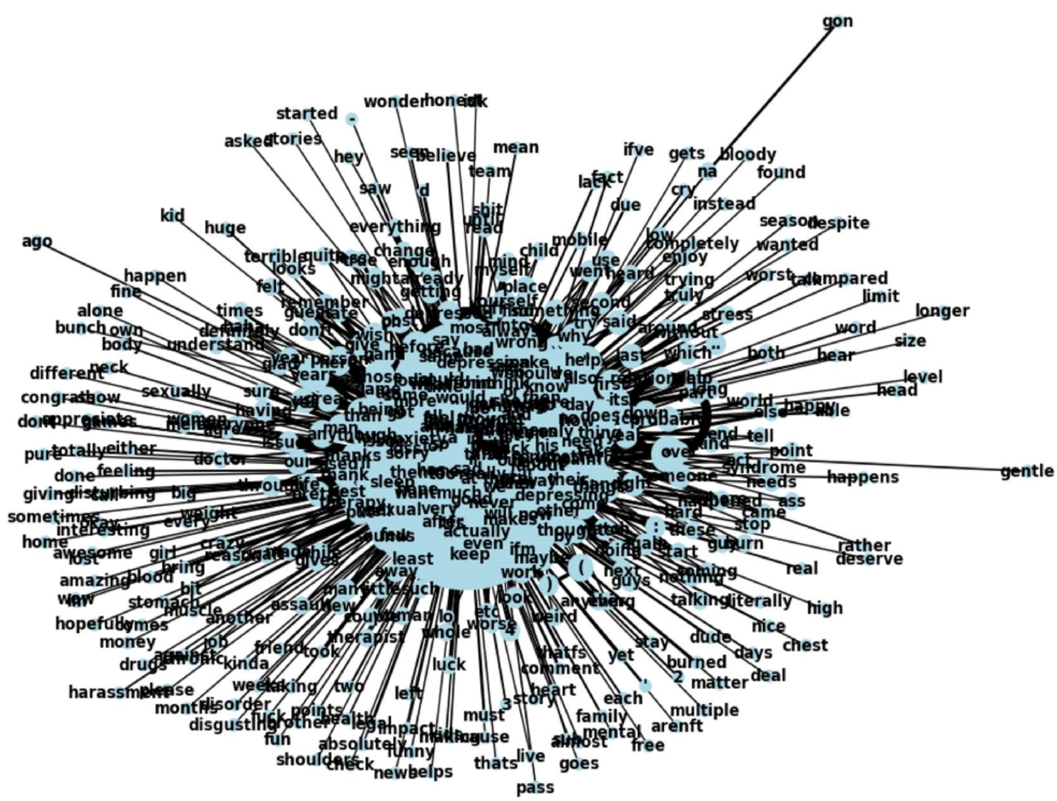

Fig. 1. Network visualization of pain-related terms. Co-occurrence network of 5,630 unique terms and 86,972 edges. Nodes represent individual terms; edges represent co-occurrence within a five-word sliding window. Node size scaled by centrality values. All nodes displayed in uniform color.

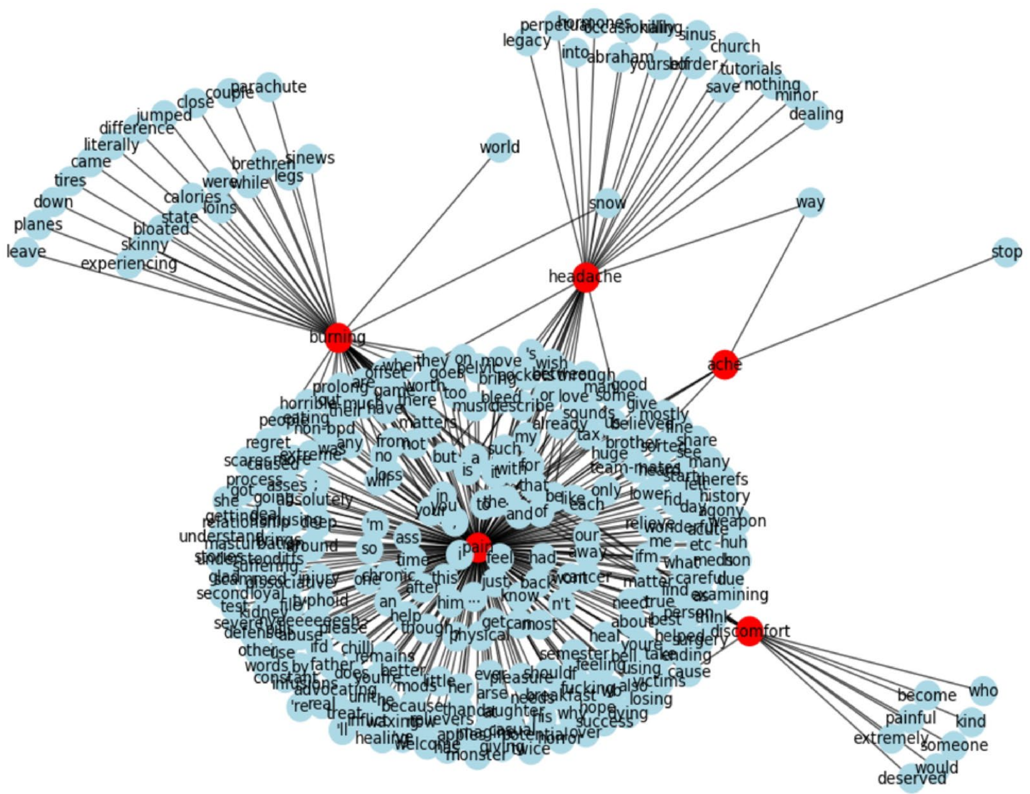

Fig. 2. Centrality measures for key pain-related terms in the network. Red nodes: primary pain-related keywords (*pain*, *headache*, *burning*, *discomfort*, *ache*). Blue nodes: secondary terms connected via co-occurrence. Node size proportional to degree centrality. Edge length inversely related to co-occurrence strength; shorter edges represent stronger associations.

0.055000), *burning* (0.182000, 0.166000, 0.110000), and *discomfort* (0.049000, 0.052000, 0.024000) displayed substantially lower scores, pointing to a pronounced structural hierarchy.

Notably, contextual analysis using the GoEmotions corpus showed that *burning* co-occurred almost exclusively with metaphorical or aesthetic descriptors (e.g., *glass*, *carving*), rather than with emotion- or symptom-oriented terms, suggesting a more figurative usage pattern.

Beyond individual terms, we examined centrality patterns across broader categories of *pain*—neuropathic, somatic, visceral, and psychosomatic—using a lexicon-based classification. Of the 21 predefined terms, only 6 (28.6%) were found in the co-occurrence network. Among them, psychosomatic terms such as *depression* and *anxiety* exhibited the highest eigenvector centralities (≈ 0.710000), indicating central placement in semantically influential regions. In contrast, clinically salient neuropathic terms like *burning* and *shooting* had lower connectivity and minimal network influence (e.g., *burning* ≈ 0.000000000000013). Somatic and visceral descriptors were largely absent, with only *pressure* marginally represented.

These findings reveal an asymmetry in how different *pain* modalities are linguistically expressed: while affective and cognitive terms dominate the discourse structure, physiologically grounded vocabulary remains peripheral or omitted. Full metrics by category are provided in Supplementary Table S1.

Statistical profiling of centrality in the pain network

This section presents a statistical characterization of centrality patterns among symptom-related terms, with the aim of evaluating structural hierarchy and connectivity within the lexical network. The *pain*-focused network exhibits a sparse but semantically ordered structure, with a density of 0.005500. The average degree is 30.900000, indicating that each node is connected to approximately 31 other terms on average. The network diameter is 5, meaning that even the most distant nodes are linked via relatively short paths, enabling efficient semantic propagation[13,14].

Centrality analysis identified *pain* as a pronounced hub node, with markedly higher values than all other terms in degree centrality (0.821429), betweenness centrality (0.930134), and eigenvector centrality (0.695893)[13,15,17]. These values indicate an integrative function across the network. Figure 3 presents a logarithmic histogram of these three centrality measures, highlighting the dominant position of *pain* in contrast to terms such as *headache* (0.107000, 0.109000, 0.055000), *burning* (0.182000, 0.166000, 0.110000), and *discomfort* (0.049000, 0.052000, 0.024000); *ache* also displayed similarly low scores (not shown in the figure).

Figure 3a shows degree centrality, which reflects the number of direct lexical connections maintained by a node. *Pain* exhibits the highest value, indicating its role in anchoring extensive semantic associations within the

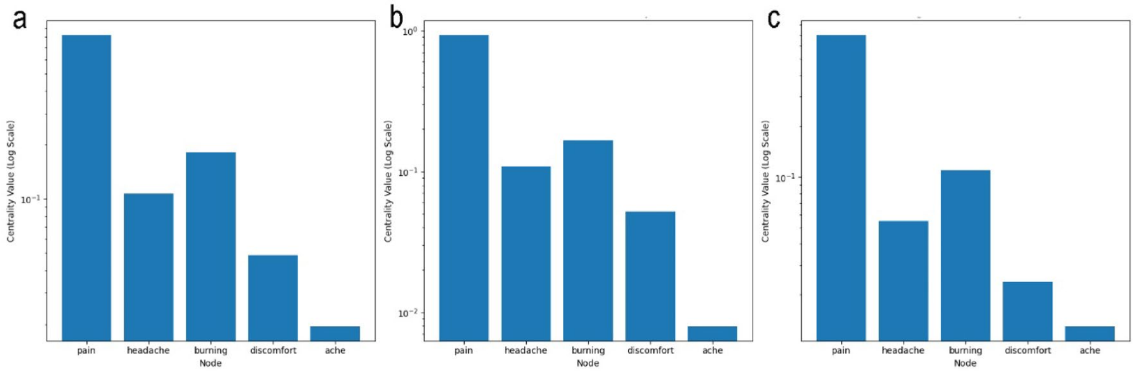

Fig. 3. Logarithmic histogram of centrality measures for key pain-related terms. (**a**) Degree centrality. (**b**) Betweenness centrality. (**c**) Eigenvector centrality. Terms: *pain*, *headache*, *burning*, *discomfort*, *ache*. X-axis: node labels. Y-axis: centrality values (log scale). Each bar represents a node.

network. In contrast, *headache* and *burning* maintain fewer but still moderate links, while *discomfort* and *ache* are sparsely connected, reflecting their marginal connectivity.

Figure 3b highlights betweenness centrality, which measures how often a node lies on the shortest paths between other nodes—indicating its function as a bridge across semantic subgroups. Here again, *pain* dominates, suggesting that it facilitates cross-cluster discourse integration. *Burning* and *headache* display lower but nontrivial betweenness values, implying partial bridging roles. Meanwhile, *discomfort* and *ache* are nearly absent from these connective pathways, underscoring their peripheral status.

Figure 3c quantifies eigenvector centrality, which evaluates a node's influence based on its proximity to other highly connected terms. The figure shows that *pain* holds a privileged position within the semantic core, exerting influence through its connections with other central terms. *Burning* ranks second in this metric, although it remains substantially lower than *pain*. In contrast, *headache*, *discomfort*, and *ache* display lower and more stable eigenvector centrality scores, indicating that they are situated in more context-specific and structurally marginal positions.

Together, these visualizations reinforce a clear structural hierarchy: while *pain* anchors the network's core, other symptom terms form secondary or peripheral nodes, playing more specialized roles within constrained lexical contexts.

Each centrality metric offers a unique interpretive lens: degree centrality reflects the number of direct connections, betweenness centrality captures the term's bridging role between semantic clusters, and eigenvector centrality quantifies influence within densely connected regions. These distinctions illuminate the functional heterogeneity of symptom vocabulary. While *pain* anchors the structure by linking diverse terms and domains, other expressions such as *headache* and *burning* occupy more context-dependent and localized roles.

Community structure in pain discourse

This section presents an analysis of how symptom-related terms cluster into semantically coherent communities, building on the centrality results described above. Based on Louvain modularity analysis, we examine whether structurally prominent terms like *pain* anchor broader thematic clusters, and how other expressions organize into distinct experiential or symbolic subdomains.

As shown in Fig. 4, the network is divided into multiple color-coded communities. The largest, labeled Community 2, contains 225 nodes and functions as a central hub of general *pain* discourse. Its average degree centrality is 0.007100, betweenness centrality is 0.004500, and eigenvector centrality is 0.046400. These values suggest that this community not only exhibits dense internal connectivity but also acts as a semantic bridge across other subgroups. With an average edge weight of 5.906200, terms in this cluster tend to co-occur frequently, reinforcing their mutual contextual associations.

The presence of additional, smaller communities reflects the thematic diversification of *pain*-related expressions. For example, Community 0, with 44 nodes and an average degree centrality of 0.008900, likely represents a moderately connected but distinct conceptual subgroup. Community 3, consisting of 14 nodes, stands out for its relatively high betweenness centrality of 0.009000, implying a mediating role despite its small size.

Close examination of the semantic profiles of specific communities reveals how *pain*-related terms are framed across experiential, emotional, and symbolic dimensions. One community, colored red and centered around the term *discomfort*, includes primarily relational and evaluative expressions such as *who*, *would*, and *deserved*. This linguistic configuration suggests that *discomfort* is often articulated in interpersonal or normative contexts, rather than through direct sensory description. Another cluster, shown in light blue and structured around *headache*, contains both symptomatic terms like *sinus* and *minor*, and systemic or metaphorical ones like *border*, *legacy*, and *outsource*. This distribution highlights the dual semantic role of *headache*, functioning both as a literal symptom and as a metaphor for social or cognitive burdens.

A third notable cluster, represented in purple and organized around *burning*, integrates terms with somatic references—such as *bloated*, *legs*, and *skinny*—alongside words associated with emotion, society, or spirituality,

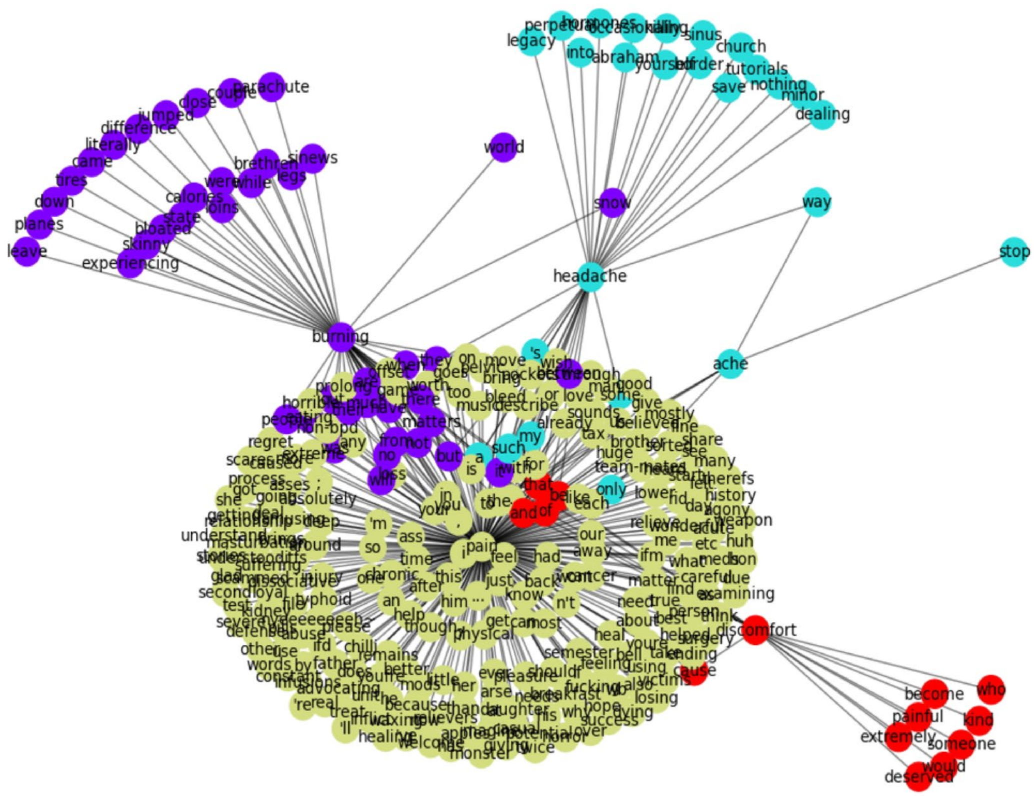

Fig. 4. Detected Communities in the Pain-Related Network Using the Louvain Method. Community structure of the pain-related lexical network, based on Louvain modularity detection. Each color denotes a distinct community. Community 0 (light blue): cluster around *headache*, including both literal symptom terms (e.g., *sinus*, *minor*) and metaphorical expressions (e.g., *border*, *legacy*). Community 1(Red): centered on *discomfort*, including relational and evaluative terms (e.g., *who*, *would*, *deserved*). Community 2 (yellow): largest cluster (225 nodes), centered on the term *pain*, representing general pain discourse. Community 3 (purple): cluster anchored by *burning*, combining somatic (e.g., *legs*, *bloated*) and symbolic terms (e.g., *regret*, *church*). Node size corresponds to degree centrality. Edge length inversely proportional to co-occurrence frequency.

including *regret*, *church*, and *world*. This blend illustrates the polysemous character of *burning*, which serves as both a physiological descriptor and a symbolic or emotional expression in discourse.

Taken together, these findings indicate that social media users construct *pain* not only as a physical sensation but also as an emotionally charged and morally situated experience. The network structure shows that linguistic representations of *pain* diverge across bodily, emotional, and social registers, yet often reconnect through central terms that bridge these domains. This pattern of differentiation and reintegration underscores the complexity of symptom discourse in online contexts.

Structural centrality of pain compared to emotion-related terms

This section presents a comparative analysis between *pain* and emotion-related terms, building on the preceding findings that established *pain*'s structural dominance and thematic centrality in the symptom discourse network. To determine whether this prominence reflects general emotional salience or a unique structural role, we examine centrality metrics of *pain* relative to two core emotion terms: *fear* and *nervousness*.

Table 1 summarizes the results of this comparison. Across all three centrality metrics—degree, betweenness, and eigenvector centrality—*pain* exhibited markedly higher values. For example, its degree centrality (0.821429) was more than six times greater than the highest value in the *fear* group (0.0937), and substantially higher than the top score in the *nervousness* group (0.1297). A similar pattern was observed in betweenness centrality (0.930134 for *pain* versus ≤ 0.2025) and eigenvector centrality (0.695893 for *pain* versus ≤ 0.3426).

Permutation tests ($n = 10{,}000$) confirmed that *pain*'s centrality scores were significantly greater than those observed in either emotion-related group ($p < 0.0001$ for all metrics). These results indicate that *pain* functions not merely as a frequent or emotionally salient term, but rather as a structurally dominant hub within the symptom discourse network. Its high connectivity and bridging role distinguish it from typical emotion terms, which tend to cluster within narrower affective contexts. This supports the interpretation of *pain* as a central organizing term that integrates diverse semantic domains.

Centrality Metric	Pain score	Max score (fear group)	Max score (nervousness Group)	Mean ± SD (combined emotion terms)	Permutation Test *p*-value (*n* = 10,000)
Degree centrality	0.821429	0.0937	0.1297	0.0557 ± 0.0291	< 0.0001
Betweenness centrality	0.930134	0.1892	0.2025	0.0921 ± 0.0374	< 0.0001
Eigenvector centrality	0.695893	0.3426	0.3267	0.1793 ± 0.0648	< 0.0001

Table 1. Centrality metrics of pain compared to fear- and nervousness-related terms. Columns: metric type (degree, betweenness, eigenvector), *pain* score, maximum scores in fear and nervousness networks, combined mean ± standard deviation for emotion-related terms, and permutation test *p*-values (*n* = 10,000). All centrality values range from 0 to 1; higher values indicate greater structural prominence.

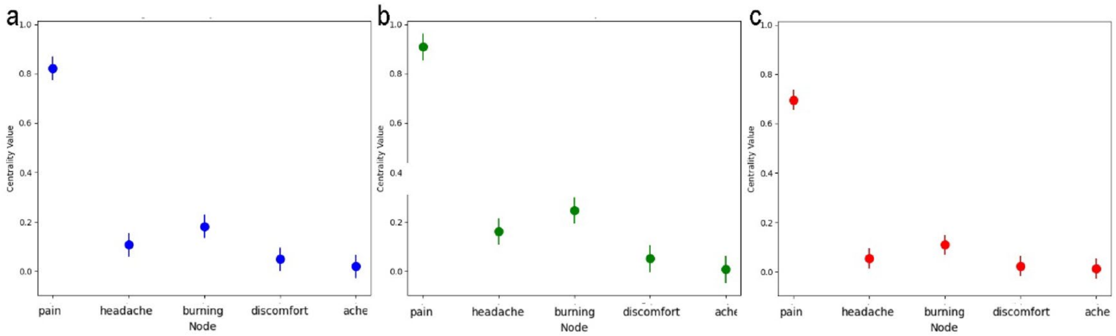

Fig. 5. Instability of centrality measures for key pain-related terms. Centrality values and standard deviations for five pain-related terms (*pain*, *headache*, *burning*, *discomfort*, *ache*). (**a**) Degree centrality. (**b**) Betweenness centrality. (**c**) Eigenvector centrality. X-axis: node labels. Y-axis: centrality values. Error bars: standard deviations.

Stability analysis of centrality measures

This section presents an evaluation of the stability of key symptom-related terms using centrality metrics and their standard deviations, building on the previous structural analysis. Figure 5 provides a comparative visualization of centrality values and their variability for five key terms—pain, headache, burning, discomfort, and ache—across three centrality measures. Each panel in the figure corresponds to a specific metric: Fig. 5a shows degree centrality, Fig. 5b shows betweenness centrality, and Fig. 5c displays eigenvector centrality. Points represent mean centrality scores, while vertical error bars indicate standard deviations calculated from bootstrap resampling.

In Fig. 5a, pain shows the highest degree centrality with a moderate standard deviation, indicating that it consistently forms a dense set of direct lexical connections across samples. Burning and headache follow with lower scores and slightly higher variability. Discomfort and ache show both low centrality and limited variation, suggesting marginal connectivity and functional specificity in the network.

Figure 5b highlights pain's role as a semantic bridge, with a betweenness centrality of 0.930134 and substantial variability, reflecting its dynamic positioning across contextual subgroups. Headache displays a standard deviation of 0.161610—greater than other non-central nodes—indicating its context-dependent bridging function. In contrast, discomfort and ache appear rarely along shortest paths between clusters, affirming their peripheral discursive role.

Figure 5c quantifies influence within densely connected regions. Pain again dominates, with burning positioned second in influence but showing a greater variance. Headache, discomfort, and ache all exhibit lower and more stable eigenvector centralities, reinforcing their limited integration into the semantic core of the network.

To further assess the robustness of these findings, 95% confidence intervals (CIs) for degree centrality were calculated using bootstrap standard deviations. Pain had the widest CI (0.193–1.450), underscoring both its dominant and context-sensitive role. Headache showed a CI of 0.020–0.195, while burning ranged from 0.034 to 0.329. Both discomfort and ache maintained lower and narrower ranges, each under 0.090, suggesting stable but structurally marginal positions.

To determine whether centrality variability differs by semantic category, we compared the standard deviations of degree centrality across three lexical groups: pain-related, metaphorical, and emotional terms. Statistical tests revealed that pain-related terms exhibited significantly higher variability than metaphorical terms (Welch's $t = 4.11$, $p = 0.0052$; permutation $p = 0.0034$), whereas no significant difference was observed between pain-related and emotional terms ($p > 0.77$). These results suggest that variability is not randomly distributed but reflects functional distinctions within the discourse structure.

Taken together, the centrality stability profiles clarify the structural resilience and contextual adaptability of symptom-related terms. Pain consistently occupies a central yet flexible role, while headache and burning function within narrower, more variable contexts. In contrast, discomfort and ache remain stable but peripheral, contributing minimally to the broader organization of symptom-related language. These findings highlight the importance of incorporating stability metrics into network-based discourse analysis.

Discussion

This study reveals a pronounced structural hierarchy in pain-related discourse, with *pain* functioning as a semantic anchor that organizes symptom expressions across diverse communicative contexts[3,5,18]. This centrality extends beyond frequency, reflecting *pain*'s role as a linguistic attractor that binds disparate sensory, emotional, and cognitive descriptors into coherent symptom narratives[3,10]. The network's sparse overall density (0.0055) is counterbalanced by a high clustering coefficient (0.7700), indicating tightly knit local groupings. Moreover, the network exhibits a short diameter of 5—that is, even the most distant terms are connected by five or fewer co-occurrence steps—demonstrating efficient semantic linkage across the lexical structure[19,20].

Within this framework, *pain* consistently dominates all centrality measures, while terms like *headache*, *burning*, and *discomfort* occupy more peripheral or context-specific positions[21,22]. This asymmetric configuration suggests that symptom discourse is not evenly distributed but shaped by key organizing principles[20]. The observed hierarchy also reflects the phenomenological reality that *pain* serves as both a descriptive label and an interpretive framework through which bodily experiences are conceptualized[5]. Unlike other sensory or emotional terms that cluster within narrower affective subdomains, *pain* demonstrates remarkable semantic breadth, linking physiological states to psychological and social meanings[3,5,10].

This structural organization holds critical implications for how individuals conceptualize and communicate suffering, positioning *pain* not merely as a symptom descriptor but as a foundational organizing concept in health-related language[3,5,10].

While centrality metrics provide valuable insight into lexical prominence, they alone cannot capture the full functional complexity of symptom-related vocabulary[21,22]. The term *burning*, for instance, displays moderate centrality in the network, yet closer examination reveals a disconnect between its structural presence and its clinical function[23,24]. Co-occurrence analysis within the GoEmotions corpus shows that *burning* frequently appears in metaphorical or decorative contexts—such as *burning glass* or *burning wood*—rather than in *pain*-related or emotional narratives[23–25].

In contrast to its clinical role as a hallmark descriptor for neuropathic pain, *burning* in digital discourse fails to maintain semantic proximity to other *pain* terms, highlighting a fundamental divergence between medical language and everyday usage[26,27]. Similarly, *headache* exhibits high variability across centrality metrics, suggesting that its discursive role shifts depending on context—sometimes used literally, other times metaphorically or emotionally[27,28]. This functional plasticity implies that such terms operate as context-sensitive bridges within the discourse, adapting their meaning based on narrative framing[29,30].

These observations underscore the need for contextual validation in network-based health communication studies[24,29]. Metrics alone cannot account for the pragmatic dimensions of lexical meaning. Instead, terms must be evaluated within the discursive ecosystems they inhabit, where semantic roles are negotiated dynamically rather than statically defined[30,31].

The pragmatic divergence observed above, particularly in the case of *burning*, exposes deeper structural issues in the clinical use of diagnostic vocabulary[31,32]. In the context of *interstitial cystitis/bladder pain syndrome*, *burning sensation* is a central diagnostic criterion embedded in standardized questionnaires[33–35]. Yet our analysis shows that *burning* is rarely used in *pain*-related contexts on social media platforms. This suggests that younger, digitally native populations may not associate their lived sensory experiences with the terminology employed in formal diagnostic instruments—raising the risk of underreporting, misrecognition, or diagnostic delay[31,36].

This gap is not isolated. In cardiovascular medicine, *chest discomfort* is widely used to characterize myocardial infarction, yet the term *discomfort* in online discourse frequently appears in psychological or environmental frames rather than somatic symptom contexts[37,38]. Similarly, *tingling*—a key term in describing neuropathic symptoms such as diabetic neuropathy or postherpetic neuralgia—is often employed in metaphorical expressions to denote emotional states[39,40].

These patterns reflect a broader generational and linguistic shift that challenges the assumptions embedded in patient-reported outcomes and symptom checklists[41,42]. As clinical vocabularies fail to align with the language patients use organically, the risk of miscommunication increases[43,44]. Bridging this semantic gap will require systematic updates to diagnostic tools, guided by empirical discourse analysis, to ensure that clinical instruments remain intelligible, resonant, and effective in contemporary health communication[45,46].

The variability observed in centrality measures should not be dismissed as methodological noise but rather interpreted as a structural indicator of functional diversity within symptom discourse[47]. This variability reflects the adaptive capacity of lexical items to operate across multiple narrative contexts, revealing fundamental differences in how terms function within the semantic ecosystem[48,49]. Terms with high centrality coupled with moderate instability, such as *pain*, demonstrate remarkable functional plasticity—maintaining structural dominance while adapting to diverse communicative situations[3]. This flexibility enables *pain* to serve as a versatile organizing principle across physiological, emotional, and social registers, explaining both its consistent prominence and contextual adaptability[5,6].

In contrast, terms like *discomfort* and *ache* exhibit low centrality with minimal variability, suggesting constrained semantic functions and stable but peripheral roles. Their consistent positioning reflects specialized usage patterns that resist contextual adaptation, indicating narrow functional niches within pain discourse[13,21]. The statistical comparison across semantic categories further illuminates this pattern: pain-related terms showed significantly higher variability than metaphorical expressions (Welch's $t = 4.11$, $p = 0.0052$), while showing no significant difference from emotional terms, suggesting that core symptom descriptors share the contextual flexibility characteristic of affective language[50].

These findings reveal a fundamental principle of lexical organization in health discourse: the coexistence of "fluid centers" and "fixed peripheries." While central terms like *pain* maintain their structural importance through adaptive flexibility, peripheral terms preserve their positions through functional specificity[17,20]. This hierarchical arrangement reflects how symptom vocabulary balances semantic stability with communicative versatility, enabling both precise description and flexible meaning negotiation[51].

Understanding the structural hierarchy and functional variability of symptom vocabulary offers significant implications for clinical communication and practice. Building on the structural insights outlined above, *pain* can be strategically leveraged as a central lexical anchor in diagnostic contexts, providing clinicians with a valuable entry point for eliciting more comprehensive symptom narratives[5,7]. The identification of stable peripheral terms versus variable bridging concepts provides a framework for interpreting patient language patterns, potentially improving triage efficiency and educational interventions by focusing on the lexical nodes through which patients naturally structure their experiences[29,46].

The divergence between clinical terminology and everyday language usage, particularly evident in terms like *burning*, suggests the need for bridging models that can translate between professional medical discourse and patient-generated descriptions. Such translation frameworks could enhance patient-provider communication by recognizing how symptom descriptors function differently across contexts, enabling more accurate interpretation of patient reports and more effective health education strategies[31,36].

However, several limitations constrain the immediate clinical applicability of these findings. In particular, social media users represent a self-selected and demographically skewed population—typically younger, digitally fluent, and culturally specific—limiting the generalizability of these findings to more diverse clinical populations. The reliance on social media data may not fully capture the linguistic patterns present in direct clinical encounters, and the temporal stability of these network structures requires validation across different time periods and demographic populations[9,24]. Moreover, while our quantitative network analysis provides structural insights into pain-related discourse, it should be interpreted alongside qualitative and ethnographic approaches that capture the lived experiences, cultural narratives, and interpersonal dynamics that shape how pain is communicated and understood[52,53]. Future research should integrate electronic medical records with patient-reported outcome measures to validate the clinical relevance of social media-derived language patterns[11,12]. Although our analyses confirmed the robustness of key network properties, we acknowledge that formal comparisons with graph-randomized or degree-preserving null models were not performed. Incorporating such null model frameworks would allow more rigorous inference of structural significance for metrics like clustering and modularity, and represents an important direction for future methodological work.

From a methodological perspective, this discrete mathematical approach to symptom language analysis provides a quantitative foundation for advancing personalized patient communication strategies in clinical practice[54,55]. The development of hybrid analytical frameworks that bridge social media linguistics with clinical communication represents a promising direction[56]. Such approaches could potentially transform network-derived centrality metrics into predictive indicators for patient communication preferences, symptom progression patterns, and treatment adherence, ultimately advancing personalized approaches to clinical dialogue and care delivery.

Methods

Ethics statement

This study analyzed publicly available data from the GoEmotions dataset, which contained anonymized Reddit comments. According to the platform's terms of service, Reddit users consent to their public posts being viewed and analyzed. No additional ethical approval was required as this study used only publicly available, anonymized data and did not involve any direct human participant interaction. All data handling complied with Reddit's terms of service and data-usage policies.

Data collection and preprocessing

An analysis was conducted on 57,000 Reddit comments from the GoEmotions dataset (2005–2019), which provides emotion-labeled social media texts[4,8,9]. Given the growing body of evidence suggesting that emotionally annotated language corpora can validly reflect underlying psychological constructs—including affective states, somatic perception, and interoceptive awareness[3,5,18,37], this dataset offers a suitable foundation for analyzing spontaneously expressed pain-related discourse. Recent studies further support the use of word embeddings and emotion-tagged corpora to infer nuanced emotional and bodily experiences from naturalistic text[3,57]. The dataset was pre-processed using a multistage approach. Initial cleaning involved the removal of toxic or offensive content through a combination of automated filtering (using predefined word lists) and manual annotation. To ensure the reliability of toxicity identification, a random subset of comments was independently reviewed by two annotators, yielding substantial inter-rater agreement (Cohen's $\kappa = 0.85$). In addition, only comments originating from subreddits with more than 10,000 posts were included to ensure data quality and adequate contextual richness.

Text preprocessing was performed using a custom-built NLP pipeline in Python 3.8. This included tokenization via the Penn Treebank tokenizer from the NLTK toolkit, followed by stop word removal and the elimination of special characters[25,27]. To improve terminological consistency, medical terms were standardized using a modified version of the Unified Medical Language System (UMLS) metathesaurus[31,36]. The accuracy of this standardization process was verified through manual inspection of a randomly sampled subset of 1,000 comments, achieving approximately 95% concordance between the output and clinical reference forms.

Pain-related comment identification and validation

To isolate pain-related discourse from the GoEmotions dataset, we applied a keyword-based filtering strategy. The pain-related keyword list was developed iteratively by two practicing physicians through interactive sessions with a large language model ChatGPT (OpenAI, San Francisco, CA, USA), with the goal of maximizing clinical relevance and semantic coverage. The final list was reviewed and refined under the supervision of an English language specialist (K.O.) to ensure terminological precision and consistency with biomedical discourse norms.

The resulting lexicon included over 90 pain-related terms encompassing urogenital, musculoskeletal, neuropathic, inflammatory, and psychosomatic categories. These terms were matched using case-insensitive substring search following tokenization. The full list used for filtering comprises:

genital pain, urinary pain, sexual dysfunction, dyspareunia, pain during sexual intercourse, pelvic pain, vaginal pain, erectile pain, testicular pain, bladder pain, reproductive organ pain, prostate pain, menstrual pain, intercourse pain, painful urination, genital discomfort, sexual pain, abdominal pain, headache, migraine, back pain, neck pain, shoulder pain, joint pain, muscle pain, chronic pain, acute pain, leg pain, foot pain, hand pain, stomach ache, toothache, sinus pain, chest pain, rib pain, arthritis pain, cramping, burning, throbbing, ache, soreness, discomfort, numbness, stiffness, tenderness, inflammation, spasm, nerve pain, perineal pain, bowel symptoms, urinary symptoms, bladder distension, dysuria, suprapubic pain, urethral burning, hesitancy, incontinence, frequency, urgency, sleep disturbance, fatigue, insomnia, depression, anxiety, hopelessness, sadness, loneliness, guilt, stress, apathy, worthlessness, isolation, lethargy, mood swings, irritability, withdrawal, appetite loss, emptiness, despair, suicidal thoughts, restlessness, helplessness, disability, posture issues, mobility limitations, lifestyle impact, backache, sciatica, bleeding, bloating, malnutrition, dehydration, anemia, weight loss, fever, bloody stools, diarrhea, mucus, complications, steroid therapy, immunosuppression, strictures.

Although the GoEmotions dataset was originally designed for emotion classification, its inclusion of naturalistic, emotionally annotated text makes it well-suited for identifying spontaneously expressed pain-related content in everyday language.

To assess the reliability of the filtering process, a random sample of 1,000 extracted comments was independently reviewed by N.O. and M.O. The two annotators demonstrated a high level of agreement in identifying pain-related content, confirming the consistency and thematic relevance of the filtered corpus.

Keyword identification and network analysis

Co-occurrence was defined using a five-word sliding window, in which any two terms appearing within the same five-word span were considered co-occurring. This method captures short-range contextual associations while maintaining semantic proximity and has been widely used in lexical network studies. This allowed lexical associations and patterns to emerge naturally from within the filtered subset. By analyzing co-occurrences among 123,840 potential word relationships, central terms such as pain, headache, discomfort, and burning surfaced as prominent nodes within the network.

This data-driven approach enabled the identification of emergent linguistic structures and thematic clusters without relying on externally imposed categories, offering an unbiased view of how pain is framed and communicated in social media.

Network analysis was implemented using NetworkX 2.5, employing a sliding window approach of size 5 (optimized through a sensitivity analysis of sizes 3–7). Co-occurrence weights were calculated using frequency-adjusted normalization to account for baseline term prevalence, with statistical significance assessed using Bonferroni-corrected Spearman's correlation coefficients ($p < 0.05$).

Advanced statistical analysis

Network analysis employed three centrality measures to characterize the term relationships: degree centrality to measure term connectivity, betweenness centrality to identify bridge terms between symptom clusters, and eigenvector centrality to assess node influence within the network. The community structure was detected using the Louvain method with a resolution parameter of 1.0, optimized through modularity maximization ($Q > 0.3$)[30,31]. These metrics provide complementary perspectives on the network structure of pain-related studies.

Validation and quality control

Network stability was validated using both internal and external approaches. Internal validation employed bootstrap analysis (1000 iterations) with 80/20 data splits, maintaining edge weight distribution stability (coefficient of variation < 15%) and cross-validation with fivefold partitioning stratified by year[32]. For external validation, network findings were cross-referenced with established clinical literature through a systematic review of pain comorbidity studies (2000–2023)[33].

Visualization and data representation

Network visualization was implemented using a modified force-directed layout algorithm in Python, with node sizes scaled logarithmically by term frequency. Edge weights were represented by a continuous color gradient from light grey (weak correlation) to black (strong correlation, $r \geq 0.7$)[34]. Term categories were distinguished by color: pain terms in red, associated symptoms in blue, and psychological terms in green, with node opacity reflecting term specificity scores. All visualizations were optimized for colorblind accessibility according to established guidelines. No comparisons with graph-randomized or null networks were performed in this study; however, network stability was assessed through bootstrap, perturbation, and permutation-based validation strategies.

Quality assurance and reproducibility

Text processing quality was verified through a manual review of a stratified random sample (10% of the corpus) by two independent reviewers, achieving high inter-rater reliability (Cohen's $\kappa = 0.88$ for term classification)[35]. Network stability was further evaluated through perturbation analysis, where up to 20% of the edges were randomly removed to assess changes in the community structure[31,36].

Comparative centrality analysis and permutation testing

To evaluate the statistical significance and variability of key node centrality, we focused on the term *pain* and its comparison with emotion-related terms such as *fear* and *nervousness*. Separate co-occurrence networks were constructed for each emotional category using identical preprocessing and windowing parameters. Centrality measures (degree, betweenness, and eigenvector) were extracted from each network.

For each metric, we computed the maximum and mean values for the emotion-related networks and compared them with those of *pain*. To assess statistical significance, permutation tests ($n = 10{,}000$) were conducted using the combined centrality distribution from emotion-related terms as the null distribution. *Pain* exhibited significantly higher centrality in all metrics ($p < 0.0001$).

In addition, to estimate the variability of *pain* and other key nodes (e.g., *headache*, *burning*), we performed a bootstrap analysis (1,000 iterations). In each iteration, 80% of the corpus was resampled with replacement, and degree centrality was recalculated using the same network parameters. The resulting 95% confidence intervals (mean $\pm$ 1.96 $\times$ SD) provided estimates of centrality stability across corpus subsets. This analysis complements the network-wide validation and clarifies how robustly specific terms maintain their structural prominence.

Statistical analysis

Network characteristics were analyzed using the centrality analysis framework implemented in Network X 2.5. This included the extraction and normalization of 11 key pain descriptors, which were analyzed for co-occurrence patterns using a five-word window. The analysis revealed 363 significant relationships among 309 unique terms, with co-occurrence strength normalized by the total term frequencies to prevent bias in high-frequency terms.

Network analysis employed three centrality measures (degree, betweenness, and eigenvector centrality) to identify the key hub terms and bridge concepts. Community detection was performed using the Louvain method (resolution parameter: 1.0), with significance established through bootstrap analysis (1000 iterations) and Bonferroni correction. The final integration stage included the cross-validation of centrality measures and sensitivity analyses for parameter stability. The results were visualized using matplotlib and seaborn libraries, with node sizes reflecting centrality values, and edge weights representing co-occurrence strength.

Data availability

Publicly available datasets were analyzed in this study. This data can be found at: https://github.com/hplisiecki/emotion_topology."

Code availability

The code used in this study is available at: https://github.com/Okuinobuo/PainAnalysisNLP/

Received: 9 November 2024; Accepted: 1 August 2025
Published online: 09 August 2025

References

1. Bradley, M. M. & Lang, P. J. Affective norms for English words (ANEW): instruction manual and affective ratings. Technical report C-2. *Univ. Florida, Gainesville* (2010).
2. Ekman, P., Davidson, R. J. & Friesen, W. V. The Duchenne smile: Emotional expression and brain physiology. II. *J. Pers. Soc. Psychol.* **58**, 342–353 (1990).
3. Jackson, J. C. et al. From text to thought: How analyzing language can advance psychological science. *Perspect. Psychol. Sci.* **17**, 805–826 (2022).
4. Saffar, A. H., Mann, T. K. & Ofoghi, B. Textual emotion detection in health: Advances and applications. *J. Biomed. Inform.* **137**, 104258 (2023).
5. Craig, K. D. Toward the social communication model of pain. In *Social and Interpersonal Dynamics in Pain* (eds Vervoort, T. et al.) 23–41 (Springer, 2018).
6. Geniusas, S. *The Phenomenology of Pain* (Ohio University Press, 2020).
7. Miglio, N. & Stanier, J. Beyond pain scales: A critical phenomenology of the expression of pain. *Front. Pain Res.* **3**, 895443 (2022).
8. Barrett, L. F., Quigley, K. S. & Hamilton, P. An active inference theory of allostasis and interoception in depression. *Philos. Trans. R. Soc. B Biol. Sci.* **371**, 20160011 (2016).
9. Kabir, M. K., Islam, M., Kabir, A. N. B., Haque, A. & Rhaman, M. K. Detection of depression severity using Bengali social media posts on mental health: Study using natural language processing techniques. *JMIR Form Res.* **6**, e36118 (2022).
10. Demszky, D., Movshovitz-Attias, D., Ko, J., Cowen, A., Nemade, G. & Ravi, S. GoEmotions: a dataset of fine-grained emotions. Preprint at https://arxiv.org/abs/2005.00547 (2020).
11. Heintzelman, N. H. et al. Longitudinal analysis of pain in patients with metastatic prostate cancer using natural language processing of medical record text. *J. Am. Med. Inform. Assoc.* **20**, 898–905 (2013).
12. Bacco, L. et al. Natural language processing in low back pain and spine diseases: A systematic review. *Front. Surg.* **9**, 957085 (2022).
13. Cardoso, F. M. et al. Effect of network topology and node centrality on trading. *Sci. Rep.* **10**, 11113 (2020).
14. Liu, M., Zou, X., Chen, J. & Ma, S. Comparative analysis of social support in online health communities using a word co-occurrence network analysis approach. *Entropy* **24**, 174 (2022).
15. Newman, M. E. J. *Networks: An Introduction* (Oxford University Press, 2010). https://doi.org/10.1093/acprof:oso/9780199206650.001.0001.

16. Blondel, V. D., Guillaume, J. L., Lambiotte, R. & Lefebvre, E. Fast unfolding of communities in large networks. *J. Stat. Mech.* **2008**, P10008 (2008).
17. Cavallaro, L., De Meo, P., Fiumara, G. & Liotta, A. On the sensitivity of centrality metrics. *PLoS ONE* **19**, e0299255 (2024).
18. Bunzli, S. et al. How do people communicate about knee osteoarthritis? A discourse analysis. *Pain Med.* **22**, 1127–1148 (2021).
19. Fernández-de-las-Peñas, C. et al. Understanding sensitization, cognitive and neuropathic associated mechanisms behind post-COVID pain: A network analysis. *Diagnostics* **12**, 1538 (2022).
20. Barabási, A.-L. *Network Science* (Cambridge University Press, 2016).
21. Oldham, S. et al. Consistency and differences between centrality measures across distinct classes of networks. *PLoS ONE* **14**, e0220061 (2019).
22. Yadav, A. A comparative analysis of centrality measures in complex networks. *Autom. Remote Control* **85**, 685–695 (2024).
23. Li, Z. et al. Temporal grading index of functional network topology predicts pain perception of patients with chronic back pain. *Front. Neurol.* **13**, 899254 (2022).
24. Wu, S. et al. Deep learning in clinical natural language processing: A methodical review. *J. Am. Med. Inform. Assoc.* **27**, 457–470 (2020).
25. Denecke, K. & Reichenpfader, D. Sentiment analysis of clinical narratives: A scoping review. *J. Biomed. Inform.* **140**, 104336 (2023).
26. Bouhassira, D. et al. Comparison of pain syndromes associated with nervous or somatic lesions and development of a new neuropathic pain diagnostic questionnaire (DN4). *Pain* **114**, 29–36 (2005).
27. Zhang, Y., Zhang, Y., Qi, P., Manning, C. D. & Langlotz, C. P. Biomedical and clinical English model packages for the Stanza Python NLP library. *J. Am. Med. Inform. Assoc.* **28**, 1892–1899 (2021).
28. Cheung, T. et al. Network analysis of depressive symptoms in Hong Kong residents during the COVID-19 pandemic. *Transl. Psychiatry.* **11**, 460 (2021).
29. Jones, P. J., Ma, R. & McNally, R. Bridge centrality: A network approach to understanding comorbidity. *Multivar. Behav. Res.* **56**, 353–367 (2019).
30. Zamani Esfahlani, F. et al. Modularity maximization as a flexible and generic framework for brain network exploratory analysis. *Neuroimage* **244**, 118607 (2021).
31. Xu, D. et al. Unified Medical Language System resources improve sieve-based generation and Bidirectional Encoder Representations from Transformers (BERT)-based ranking for concept normalization. *J. Am. Med. Inform. Assoc.* **27**, 1510–1519 (2020).
32. Neal, Z. P. How strong is strong? The challenge of interpreting network edge weights. *PLoS ONE* **19**, e0311614 (2024).
33. Doiron, R. C., Nickel, J. C. & Siemens, D. R. Diagnosis and management of interstitial cystitis/bladder pain syndrome: CUA guideline. *Can. Urol. Assoc. J.* **19**, 92–102 (2025).
34. Yu, W.-R., Jiang, Y.-H., Jhang, J.-F. & Kuo, H.-C. Bladder pain syndrome associated with interstitial cystitis: Recent research and treatment options. *Curr. Bladder Dysfunct. Rep.* **18**, 389–400 (2023).
35. Clemens, J. Q., Erickson, D. R. & Varela, N. P. Diagnosis and treatment of interstitial cystitis/bladder pain syndrome: AUA guideline amendment. *J. Urol.* **208**, 34–42 (2022).
36. Zheng, L. et al. A review of auditing techniques for the Unified Medical Language System. *J. Am. Med. Inform. Assoc.* **27**, 1625–1638 (2020).
37. Altstidl, J. et al. Absence of chest discomfort in type 1 NSTEMI patients: Predictors and impact on outcome. *Clin. Res. Cardiol.* **114**, 1234–1245 (2025).
38. Kumar, A. et al. Chest pain symptoms during myocardial infarction in patients with and without diabetes: A systematic review and meta-analysis. *Heart* **109**, 1516–1524 (2023).
39. Baba, M., Kuroha, M., Wasaki, Y. & Ohwada, S. Effects of mirogabalin on tingling or pins & needles in a phase 3 study of diabetic peripheral neuropathy. *J. Jpn. Soc. Pain Clin.* **27**, 287–295 (2020).
40. Wu, S., Wahle, J. P. & Mohammad, S. The Language of Interoception: Examining Embodiment and Emotion Through a Corpus of Body Part Mentions. arXiv:2505.16189 (2025).
41. Borbjerg, M. K. et al. Understanding the impact of diabetic peripheral neuropathy and neuropathic pain on quality of life and mental health in 6,960 people with diabetes. *Diabetes Care* **48**, 588–595 (2025).
42. Tofthagen, C., Visovsky, C., Dominic, S. & McMillan, S. Neuropathic symptoms, physical and emotional well-being, and quality of life at the end of life. *Support. Care Cancer* **27**, 3357–3364 (2019).
43. Hu, Y., Keloth, V. K., Raja, K., Chen, Y. & Xu, H. Towards precise PICO extraction from abstracts of randomized controlled trials using a section-specific learning approach. *Bioinformatics* **39**, btad542 (2023).
44. Engels, G. et al. Clinical pain and functional network topology in Parkinson's disease: A resting-state fMRI study. *J. Neural Transm.* **125**, 1449–1459 (2018).
45. Brusco, M. J., Steinley, D. & Watts, A. L. On maximization of the modularity index in network psychometrics. *Behav. Res. Methods.* **55**, 3549–3565 (2023).
46. Hoffman, M., Steinley, D., Gates, K. M., Prinstein, M. J. & Brusco, M. J. Detecting clusters/communities in social networks. *Multivar. Behav. Res.* **53**, 57–73 (2018).
47. Filosi, M., Visintainer, R., Riccadonna, S., Jurman, G. & Furlanello, C. Stability indicators in network reconstruction. *PLoS ONE* **9**, e89815 (2014).
48. Mokhtari, F., Akhlaghi, M. I., Simpson, S. L., Wu, G. & Laurienti, P. J. Sliding window correlation analysis: Modulating window shape for dynamic brain connectivity in resting state. *Neuroimage* **189**, 655–666 (2019).
49. Shakil, S., Lee, C. H. & Keilholz, S. D. Evaluation of sliding window correlation performance for characterizing dynamic functional connectivity and brain states. *Neuroimage* **133**, 111–128 (2016).
50. Daly, C. H. et al. Empirical evaluation of SUCRA-based treatment ranks in network meta-analysis: Quantifying robustness using Cohen's kappa. *BMJ Open* **9**, e024625 (2019).
51. Wongpakaran, N., Wongpakaran, T., Wedding, D. & Gwet, K. L. A comparison of Cohen's Kappa and Gwet's AC1 when calculating inter-rater reliability coefficients: A study conducted with personality disorder samples. *BMC Med. Res. Methodol.* **13**, 61 (2013).
52. Kleinman, A. The Illness Narratives: Suffering, Healing, and the Human Condition. (Basic Books, 1988).
53. Good, B. J. *Medicine, Rationality, and Experience: An Anthropological Perspective* (Cambridge University Press, 1994).
54. Okui, N. Laser treatment for urinary incontinence in elite female athletes analyzed using a discrete mathematics approach. *Sci. Rep.* **15**, 15450 (2025).
55. Okui, N. Innovative decision making tools using discrete mathematics for stress urinary incontinence treatment. *Sci. Rep.* **14**, 9900 (2024).
56. Turner, R. J., Hagoort, K., Meijer, R. J., Coenen, F. & Scheepers, F. E. Bayesian network analysis of antidepressant treatment trajectories. *Sci. Rep.* **13**, 8428 (2023).
57. Plisiecki, H. & Sobieszek, A. Emotion topology: extracting fundamental components of emotions from text using word embeddings. *Front. Psychol.* **15**, 1401084 (2024).

Acknowledgements

We thank Karen Okui for English language editing and proofreading of the manuscript. We are also grateful to Dr. Machiko Okui for her support in the identification and refinement of pain-related lexical terms. This paper is dedicated to the memory of my late high school friend, mathematician Yasushi Kondoh. Our shared passion for mathematics during those formative years has remained a lasting source of inspiration, ultimately leading to the writing of this work more than four decades later.

Author contributions

N.O. and S.H. contributed equally to all aspects of this research including: conceptualization, data curation, formal analysis, investigation, writing the manuscript, reviewing and editing. Both authors reviewed and approved the final manuscript.

Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Competing interests

The authors declare no competing interests.

Generative AI and AI-assisted technologies in the writing process

We used Python libraries with machine learning capabilities for statistical analysis and network visualization. However, no generative AI or AI-assisted technologies were used in the writing or editing of the manuscript text.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-025-14680-y.

Correspondence and requests for materials should be addressed to N.O.