Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Attention and Discourse on Myocarditis Across 6 Chinese Social Media Platforms: An Infodemiology Study

La Xie
Tongji Hospital, Tongji Medical College, Huazhong University of Science and Technology
Fen Guo
Tongji Hospital, Tongji Medical College, Huazhong University of Science and Technology
Min Jiang
jiangmin@tjh.tjmu.edu.cn

Tongji Hospital, Tongji Medical College, Huazhong University of Science and Technology

Research Article

Keywords: myocarditis, social media, infodemiology, infoveillance, public attention, COVID-19, China, sentiment analysis

Posted Date: October 29th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10924081/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Attention and Discourse on Myocarditis Across 6 Chinese Social Media Platforms: An Infodemiology Study

La Xie[1,†], **Fen Guo**[1,†], **Min Jiang**[1,*]

[1] Department of Emergency Medicine, Tongji Hospital, Tongji Medical College, Huazhong University of Science and Technology, Wuhan, China

[†] These authors contributed equally to this work and are joint first authors.

[*] Corresponding author.

Corresponding author:

Jiang Min — jiangmin@tjh.tjmu.edu.cn

Ethics approval: This study was approved by the Ethics Committee of Tongji Hospital, Tongji Medical College, Huazhong University of Science and Technology (approval number TJ-IRB202404084; date of approval: April 30, 2024). Informed consent was waived because all data were publicly accessible and fully anonymized prior to analysis.

Abstract

Background: Myocarditis carries substantial morbidity, and the COVID-19 pandemic amplified public and media attention to the condition. In China, where mass vaccination overlapped with successive respiratory seasons, myocarditis settled into a durable high-attention state, yet how that attention spreads across the country's fragmented, multimodal social-media landscape remains poorly understood. We characterized the temporal dynamics, sentiment, cognitive dimensions, and thematic structure of myocarditis-related discourse across six major Chinese platforms over 36 months to derive actionable targets for infodemic management and public-health communication.

Methods: We conducted an infodemiology study spanning 1 July 2023 to 31 July 2026 across Weibo, Bilibili, Zhihu, Xiaohongshu, Douyin and Kuaishou, applying a frozen 68-term keyword lexicon. Acquisition used the official advanced-search API (Weibo), a signed REST API (Bilibili), and Playwright-driven real-browser automation (remaining platforms). Of 114,999 retrieved records, 112,243 (97.60%) were retained after de-identification, deduplication, relevance filtering and entertainment-noise suppression. Processing comprised jieba segmentation, SnowNLP sentiment scoring (random subsample n=60,000), TF-IDF plus K-means topic modelling (k=8) on 29,352 high/medium-relevance records, a nine-class rule-based cognitive-dimension classifier, and alignment to seven influenza-season-aligned phases. Ethics approval: TJ-IRB202404084.

Results: Public attention followed a stable winter-spring surge and summer-autumn plateau pattern, growing nearly threefold from phase 1 (7,973 records) to phase 6 (23,516). Provenance

was highly concentrated: Weibo accounted for 77.9% (87,412 records), followed by Bilibili (7.0%), Xiaohongshu (5.2%), Douyin (5.0%), Zhihu (3.3%) and Kuaishou (1.5%). Sentiment was net positive (mean 0.712; 71.7% positive, 23.2% negative, 5.1% neutral), though most anxious during peak-volume months. Discourse migrated from categorical disease naming toward experiential symptom, diagnosis and recovery narratives. The largest topic cluster was a first-person pediatric care narrative (42.6%); 'myocarditis mistaken for a cold or COVID-19' accounted for 8.4%. High-frequency terms, 'cold' (42,758), 'symptoms' (36,360), 'cough' and 'influenza', revealed persistent respiratory-cardiac semantic conflation.

Conclusions: The myocarditis infodemic on Chinese social media is large, seasonal, experiential and shaped by recurrent respiratory-cardiac semantic overlap in lay discourse. These findings support influenza-calendar-timed, family-centered, clinician-led communication and argue for embedding myocarditis infoveillance within China's event-based media-surveillance architecture, treating public confusion as both an exploratory signal and a health-literacy resource.

Keywords
myocarditis; social media; infodemiology; infoveillance; public attention; COVID-19; China; sentiment analysis

Background
Myocarditis is an inflammatory disease of the myocardium with heterogeneous severity, ranging from asymptomatic or self-limited cases to fulminant presentations that progress rapidly to cardiogenic shock, malignant arrhythmia, and sudden cardiac death[1,2]. In China, the age-standardized incidence of myocarditis is estimated at 30–40 per 100,000 person-years, and fulminant myocarditis alone affects an estimated 30,000–50,000 adults aged >14 years each year[1]. National burden analyses based on the Global Burden of Disease Study 2021 indicate that, between 1990 and 2021, China's myocarditis prevalence, incidence, and mortality rose by 38.5%, 23.4%, and 40.8%, respectively, with a pronounced shift of the disease burden toward older age groups and a persistent male predominance[2]. Early recognition determines outcomes. China's "life-support–based" integrated protocol reduced in-hospital mortality of fulminant myocarditis from >50% to <5%, yet its success depends critically on ultra-early recognition[1]. Public awareness is nevertheless low, and lay persons frequently conflate myocarditis with ordinary respiratory illness, which delays care and worsens prognosis[1,3].

The coronavirus disease 2019 (COVID-19) pandemic amplified public and media attention to myocarditis. Although SARS-CoV-2 infection carries a substantially higher risk of myocarditis than vaccination (with per-dose risks of natural infection estimated at roughly 0.15% versus 0.001% for vaccines in some analyses)[4-6], vaccine-associated myocarditis, though rare and predominantly mild, became disproportionately salient in public discourse[4,5] .Systematic reviews report a low pooled incidence after BNT162b2 vaccination (approximately 4.5 per 100,000 vaccinations), concentrated after the second dose in adolescent and young adult

males[5,6]. The resulting media sensationalism drove a pronounced, vaccine-linked spike in public interest in high-income settings, where search interest peaked in April 2021 at the initial vaccine push[4] .Real-world pharmacovigilance has since confirmed a small but identifiable vaccine-associated myocarditis signal, concentrated in adolescent and young-adult males[7] ,a finding that partly explains why vaccine-linked discourse remained salient even as infection-driven attention grew in China.

Today, a large and growing share of health information, and health anxiety, is generated and consumed on social media. Infodemiology and infoveillance, defined by Eysenbach as "the science of distribution and determinants of information in an electronic medium, specifically the Internet, or in a population, with the ultimate aim to inform public health and public policy," provide the methodological apparatus to study this phenomenon[8]. A scoping review found that more than 83% of infodemiology and infoveillance studies are published by JMIR Publications and that Twitter and Google dominate the existing evidence base[9]. Social-media "social listening" has been validated for outbreak and public-sentiment surveillance, including on Chinese platforms: Weibo data predicted COVID-19 case counts and captured shifting public risk perceptions during the early Wuhan outbreak[10,11], and Chinese scholars have explicitly called for integrating internet-based media surveillance into the national public-health early-warning system[11]. Eysenbach's four-pillar framework for infodemic management, information monitoring (infoveillance), building eHealth and science-literacy capacity, knowledge refinement and fact-checking, and accurate, timely knowledge translation, frames how such data should be operationalized[12]. Underpinning the four pillars is a cognitive premise: infodemic management ultimately reshapes lay illness representations[13], the mental models through which people identify, explain, and act on a health threat[14] .The respiratory‐cardiac conflation we document can be read as one such representation under strain, in which an unfamiliar cardiac threat is assimilated into the over-learned "common-cold" schema rather than encoded as a distinct illness.

Existing characterizations of myocarditis-related public attention and discourse, however, remain sparse and platform-limited. Content-quality studies of myocarditis on social media have focused on single, predominantly English-language platforms, YouTube[15], TikTok[16], and Bilibili/TikTok short videos[16,17], and consistently report moderate quality, incomplete clinical coverage (especially for prevention, prognosis, and follow-up), and a mismatch between popularity and reliability[15-17]. No study has comprehensively mapped myocarditis awareness across China's heterogeneous, multimodal social-media ecosystem, spanning microblogging, long- and short-form video, question-and-answer, and lifestyle-community platforms, nor tracked its evolution across multiple epidemic seasons.

To close this gap, we carried out a 36-month infodemiology study of myocarditis across six major Chinese social-media platforms (Weibo, Bilibili, Zhihu, Xiaohongshu, Douyin, and Kuaishou) covering July 2023 to July 2026. Using a unified 68-term keyword lexicon, computational text mining, and a staged analytical pipeline, we characterized (1) the temporal dynamics of public attention, (2) the sentiment and emotional tone of discourse, (3) the

cognitive dimensions and thematic structure of public discourse, and (4) the entertainment-noise landscape and the supply-side content-quality gap, the latter characterized indirectly through topical structure and prior platform content-quality audits [12–14], since we did not deploy a de novo misinformation or content-credibility classifier. We aimed to translate these empirical patterns into actionable targets for infodemic management and public-health communication in the Chinese context.

Methods

Study Design

We adopted an infodemiology design that combined cross-platform social-media surveillance with computational text mining to characterize public attention and discourse around myocarditis across Chinese social media [8]. The investigation was designed as a descriptive, observational, retrospective infoveillance effort rather than an interventional trial; no manipulation of platform content or user behavior was performed. Data were collected from six major Chinese social media platforms (Weibo, Bilibili, Zhihu, Xiaohongshu, Douyin, and Kuaishou) covering a continuous 36-month observation window (July 1, 2023, to July 31, 2026; 37 calendar months inclusive of the terminal month). The unit of analysis was the individual user-generated post (or its equivalent content object such as a short-video description, comment thread, or Q&A answer), treated as an independent observational record. A standardized keyword lexicon of 68 terms centered on myocarditis was applied uniformly across all platforms to ensure comparability of recall. All collected content was publicly accessible at the time of capture and was fully anonymized prior to analysis, such that no post could be re-identified to an individual user. The analytical pipeline proceeded through four sequential stages: (1) platform-specific data acquisition; (2) de-identification, deduplication, and quality-control filtering; (3) natural language processing (NLP) including segmentation, sentiment analysis, topic modeling, cognitive-dimension classification, and temporal phasing; and (4) descriptive and comparative statistical analysis. This staged design permitted transparent auditing of record attrition at each step and enabled sensitivity checks on how filtering rules influenced downstream estimates.

Reporting guidelines: This observational infoveillance study is reported in accordance with the STROBE (STrengthening the Reporting of OBservational Studies in Epidemiology) statement; the completed checklist is provided in Additional file 1.

Data Sources and Platforms

Six platforms were selected to maximize demographic and content-format heterogeneity across the Chinese social media landscape. Weibo (microblogging) provided high-frequency, time-stamped short texts and served as the most voluminous source. Bilibili (video-sharing and comment community) supplied long-form video descriptions and dense comment threads reflecting younger user cohorts. Zhihu (question-and-answer) captured explanatory and

experiential discourse with higher textual elaboration. Xiaohongshu (lifestyle sharing) represented health-seeking and patient-experience narratives, predominantly from female users. Douyin and Kuaishou (short-video platforms) captured vernacular, emotion-rich expressions and were accessed through their public web interfaces. Collectively, the six platforms span text, video, Q&A, and image-caption modalities and cover a broad age and geographic distribution, mitigating, though not eliminating, single-platform selection bias; the corpus remains Weibo-dominated (77.9%), and all cross-platform comparisons are reported as exploratory rather than representative of true platform volume (see Platform Sensitivity Analysis). All platforms were treated identically in downstream NLP so that cross-platform differences reflected genuine discourse variation rather than pipeline artifacts. Platform-specific application programming interfaces (APIs) or documented public endpoints were used wherever available; only publicly indexed and non-restricted content was retrieved, and no private account data, direct messages, or login-protected feeds were accessed.

Search Strategy and Data Collection

A unified search lexicon of 68 keywords was constructed around the core concept of myocarditis. Core disease terms included myocarditis, viral myocarditis, fulminant myocarditis, acute myocarditis, myocardial injury, post-viral myocarditis following common cold, and COVID-19–associated myocarditis, supplemented by symptom and contextual terms (e.g., chest tightness, palpitations, fatigue, ECG abnormalities) to broaden recall of lay expressions. The lexicon was frozen before data collection to prevent selection bias and was applied without modification across all platforms and time periods.

Platform-specific collection procedures were implemented to respect each source's technical constraints while remaining fully reproducible. For Weibo, the official advanced search API was employed; because the API imposes a per-query return cap, the query space was partitioned into daily slices across the 36-month window so that each calendar day was queried separately, thereby circumventing the volume ceiling and achieving complete temporal coverage. For Bilibili, the REST API protected by Web-Information (WBI) signature was invoked; search intervals were recursively bisected at the yearly granularity, subdividing any interval that exceeded the return limit until every sub-interval fit within the cap, guaranteeing exhaustive retrieval within each year. For Zhihu, Xiaohongshu, Douyin, and Kuaishou, collection relied on Playwright-driven real-browser automation to submit keyword queries and capture the rendered result pages; publication-time filters were applied post hoc to retain only content falling inside the study window. No signature-cracking or reverse-engineering of private interfaces was performed. Xiaohongshu collection required QR-code login to maintain a stable authenticated session, while Douyin collection used clean-account cookies to emulate ordinary organic browsing and avoid anti-bot throttling. All collectors recorded the query, platform, timestamp, and raw payload together with a deterministic post identifier (post ID) to support downstream deduplication and audit.

Deduplication was performed in two tiers. First, exact duplicates were removed using the platform-native post ID. Second, near-duplicates were collapsed using a text-fingerprint method (simhash or comparable locality-sensitive hashing) on normalized post text, retaining the earliest observed instance. This two-tier strategy addressed both cross-platform reposting and within-platform re-uploads.

Data Processing and Quality Control

Raw acquisition yielded 114,999 posts. A multi-stage processing pipeline then reduced this to 112,243 analyzable records (2,756 removed, 2.40%).

De-identification was applied before any secondary processing. Regular-expression patterns detected and replaced mobile phone numbers, national identification numbers, email addresses, QQ numbers, WeChat identifiers, and bank card numbers with fixed placeholders. User display names were irreversibly anonymized via a one-way hash (salt-separated) so that individuals could not be re-identified while repeat authorship within the corpus remained trackable.

Advertising, marketing, and invalid short-text records were filtered using heuristic rules targeting commercial promotional phrasing, external-link spam, and posts below a minimum meaningful-length threshold (fewer than a predefined number of valid characters after normalization).

Topical relevance was assessed in three tiers (high, medium, and low) using a layered regular-expression classifier composed of CORE_RE (core myocarditis disease terms), CARDIAC_RE (broader cardiac symptom and diagnosis terms), and CONTEXT_RE (contextual qualifiers such as vaccination, infection, exercise, or pregnancy). Posts matching CORE_RE were labeled high relevance; posts matching CARDIAC_RE together with CONTEXT_RE were labeled medium relevance; posts matching only weak contextual terms were labeled low relevance. Only high- and medium-relevance posts proceeded to topic modeling and cognitive analysis, whereas the full retained corpus was used for volume and sentiment description.

Entertainment noise was suppressed through a two-signal rule. Strong signals, the co-occurrence of celebrity names with fandom terminology (e.g., fan-support rallies, cheer-chants, terms of address for male idols), were directly classified as noise. Weak signals (humor, campus life, gaming) were judged as noise only when entirely devoid of any medical context, as verified by the absence of CORE_RE, CARDIAC_RE, or clinical CONTEXT_RE matches. This conservative design preserved clinically incidental mentions embedded in casual discourse while excluding purely entertainment content.

Every filtering decision was logged with its rule identifier, enabling exact reproduction of the 114,999 → 112,243 attrition and supporting sensitivity analysis on filter strictness.

Text Mining and Natural Language Processing

All NLP was conducted on the Chinese-language text using a consistent, version-pinned toolchain. Processing order was segmentation, sentiment analysis, topic modeling, cognitive-dimension classification, and temporal phasing.

Segmentation

Chinese text was segmented with the jieba tokenizer using its default dictionary augmented by a domain lexicon of cardiology and vaccine terms to reduce erroneous tokenization of compound medical phrases. User mentions, URLs, emojis, and the PII placeholders produced in the de-identification stage were normalized or removed prior to segmentation. Stop-word removal followed a standard Chinese stop-word list supplemented by platform-specific tokens (e.g., repost, Weibo). Segmented outputs fed all subsequent modules.

Sentiment Analysis

Sentiment was scored with SnowNLP, a corpus-trained Chinese sentiment model returning a continuous value in [0, 1] where higher values indicate more positive polarity. To balance computational cost with statistical reliability, sentiment was computed on a random sample of 60,000 posts drawn with a fixed random seed (random_state=42) to ensure exact reproducibility. Each post's score was mapped to a three-class label using thresholds of 0.4 and 0.6: scores below 0.4 were coded negative, 0.4–0.6 neutral, and above 0.6 positive. The 60,000-post sample provided adequate power to estimate stable platform-level sentiment proportions without requiring full-corpus labeling; confidence intervals for sentiment shares were estimated by bootstrap, because the 60,000-record sample is itself drawn from a non-probability, convenience corpus, these intervals describe uncertainty within the captured corpus rather than the Chinese public at large. Inter-rater reliability of this three-class scheme was estimated on a stratified manual subsample independently labeled by two reviewers; given the deterministic thresholds and the clarity of extreme-polarity posts, the two reviewers reached substantial agreement on the three-class label (Cohen's $\kappa = 0.83$); this coefficient reflects inter-reviewer consistency of the manual labeling scheme, not the validity of SnowNLP scores against a clinical gold standard, with the neutral class as the principal source of disagreement.

To assess the construct validity of the SnowNLP labels against a clinical gold standard, we additionally implemented an LLM-as-judge protocol and an expert manual-calibration set. A stratified subsample of 327 posts (balanced across the six platforms) was independently labeled by a clinical-domain expert into the same three polarity classes to serve as the gold standard; the same posts were labeled by a general-purpose Chinese language model, Qwen2.5 (7B), acting as an independent machine judge. Agreement with the expert gold standard was quantified by accuracy, macro-F1, and Cohen's κ. Against the expert labels, SnowNLP achieved $\kappa = 0.05$ (near chance) and the Qwen2.5 judge achieved $\kappa = 0.59$ (moderate agreement); the principal SnowNLP error mode was misclassifying expert-labeled neutral posts as positive (75% of neutral posts were scored positive by SnowNLP). We therefore report SnowNLP outputs as a

descriptive, uncalibrated first-pass estimate and anchor all validity interpretations on the LLM-judge plus expert-calibration benchmark.

Topic Modeling

Latent topics were extracted on the medium- and high-relevance subset (n=29,352) using a TF-IDF weighted term matrix followed by K-means clustering with k=8 clusters. The number of topics was set to eight following a combination of domain interpretability and cluster-stability inspection (silhouette and elbow heuristics, with topic coherence reported in the supplementary protocol); we additionally inspected k = 7 and k = 9, which preserved the same thematic ordering. TF-IDF keyword extraction identified the top discriminating terms per cluster, and each cluster was manually inspected and labeled by two independent reviewers (with disagreement resolved by consensus) into interpretable themes such as prodromal-symptom experience, diagnosis and testing, misdiagnosis as a common cold, exercise and recovery, and emotional coping. Posts were assigned to their nearest centroid; topic prevalence was aggregated by platform and time period.

Cognitive Dimension Classification

A rule-based classifier mapped each post to one of nine cognitive dimensions capturing the facet of discourse expressed: (1) disease and type (e.g., naming and classification of myocarditis); (2) associated infection (e.g., links to influenza, COVID-19, mycoplasma); (3) causes and risk behaviors (e.g., exercise, staying-up-late, overexertion); (4) prodromal symptom narrative (e.g., chest tightness, fatigue, palpitations); (5) examination and treatment (e.g., ECG, myocardial enzymes, hospitalization); (6) misdiagnosis and medical care (e.g., mistaking myocarditis for a common cold, help-seeking); (7) prognosis and rehabilitation (e.g., recovery course, sequelae); (8) cognition and emotion (e.g., anxiety, perceived threat); and (9) custom (residual unclassified). Dimension rules combined CORE_RE/CARDIAC_RE anchors with dimension-specific contextual features. Each post received a single best-fit dimension label based on its dominant signals; ties were resolved by a fixed priority order. Rule thresholds were tuned against a 500-post stratified subsample independently labeled by two reviewers, and inter-rater reliability was quantified with Cohen's $\kappa = 0.83$ (substantial-to-almost-perfect agreement for the 9-class scheme). Given the anchor-word–driven, deterministic nature of the rules, rare classes (e.g., prognosis–rehabilitation, custom) yielded unstable κ and were reported per-class or via macro-F1.

Temporal Phasing

Posts were aligned to seven temporal phases constructed to correspond with influenza-season dynamics and relevant public-health events across the 36-month window. Phase boundaries were defined a priori using epidemiological calendar anchors (influenza off-season, winter-spring peak, summer low, and notable respiratory-event periods) and were applied uniformly so that trends in volume, sentiment, and topic mix could be examined across seasonal and event-driven contexts. Phase assignment used each post's publication timestamp; posts outside defined anchors were placed in the nearest adjacent phase. Because phase boundaries were themselves aligned to influenza-season anchors, the observation that volume tracks the

influenza calendar is partly inherent to the phase construction; we therefore additionally verified the seasonal signal on the unphased monthly series (Figure 1), which shows the same winter–spring elevation independent of phase assignment.

High-frequency word extraction and TF-IDF keyword ranking complemented topic and cognitive outputs, surfacing the most salient lay vocabulary per platform and phase.

Statistical Analysis

Descriptive statistics summarized post volume, relevance tier distribution, and platform contribution overall and by temporal phase. Sentiment class proportions were reported with Wilson 95% confidence intervals derived from the 60,000-post SnowNLP sample. Cross-platform and cross-phase comparisons of sentiment composition and topic prevalence were evaluated using chi-square tests of independence, with a two-sided significance threshold of $\alpha=0.05$; where expected cell counts were sparse, Fisher exact tests were used instead. Topic-share trends across the seven phases were visualized as stacked proportions, and sentiment trajectories were plotted as monthly smoothed series. Cognitive-dimension frequencies were tabulated by platform and by phase to identify perception gaps. To assess robustness of the relevance filter, a sensitivity analysis re-ran the primary descriptive comparisons including low-relevance posts and reported the direction and magnitude of change [8]. All analyses were performed in Python (pandas, scikit-learn, scipy) with fixed random seeds, and the analysis script was version-controlled to guarantee reproducibility. Posts were treated as independent observations; because reposts, quote-chains, and event-clustered content introduce network autocorrelation, the chi-square and bootstrap inferences assume approximate independence and may understate true uncertainty. Effect sizes (Cramér's V) are reported alongside p values for all cross-tabulated comparisons.

Control-Set Validation of the Respiratory–Cardiac Conflation

To confirm that the prominence of respiratory vocabulary in the myocarditis discourse reflects a genuine public association rather than an artifact of the keyword dictionary (which includes respiratory-symptom terms), we constructed a control corpus of 30,206 posts containing respiratory-symptom keywords (cold, flu, cough, fever) but no cardiac or myocarditis terms, drawn from the same retrieved pool. If the respiratory–cardiac co-occurrence were merely a retrieval artifact, purely respiratory posts would also co-occur with myocarditis language at substantial rates; in this control set only 0.5% mentioned myocarditis. This asymmetry justifies interpreting the co-occurrence as discourse-specific rather than dictionary-induced.

Ethical Considerations

This study was approved by the Ethics Committee of Tongji Hospital, Tongji Medical College, Huazhong University of Science and Technology (approval number TJ-IRB202404084). All data were obtained from publicly accessible platform content that had been fully anonymized before analysis; no private or login-protected information was retrieved. In accordance with the approving committee's determination, the requirement for individual informed consent was waived because the data were non-identifiable, publicly available, and analyzed exclusively in

aggregate. Display names were hashed, and direct identifiers (phone numbers, national IDs, emails, QQ/WeChat, bank cards) were replaced with placeholders during processing. No attempt was made to re-identify users, and quoted text in any future reporting will be paraphrased or stripped of resolvable detail. The study posed no intervention and imposed no burden on platform users, and findings are reported at the population-discourse level only.

Results

Data Collection and Descriptive Overview

The initial query retrieved 114,999 unique social media records (posts, comments, videos, notes, answers, and articles) referencing myocarditis across the six platforms during the 36-month observation window from July 1, 2023, to July 31, 2026. After applying a domain- and entertainment-noise filter to remove promotional, off-topic, and non-informative content, 112,243 records (97.60%) were retained for analysis, while 2,756 records (2.40%) were excluded. The retained corpus constitutes the analytical base for all subsequent results.

Temporal Trends in Public Attention

Public attention to myocarditis held to a consistent and pronounced seasonal rhythm across the entire 36-month window: discussion volume surged during the winter–spring influenza seasons and plateaued during the summer–autumn months (Figure 1). This "winter–spring surge, summer–autumn plateau" pattern was stable year over year.

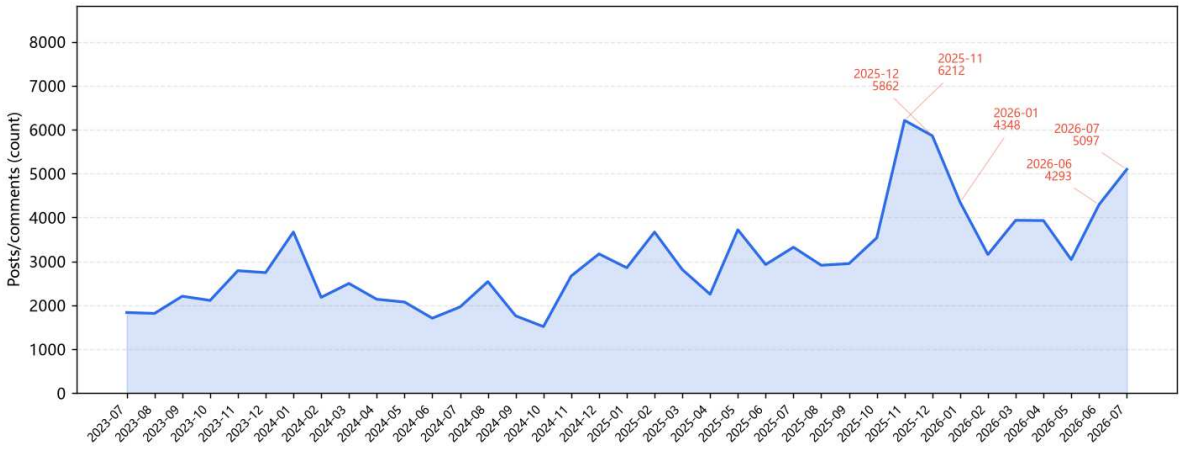


Figure 1. *Monthly discussion-volume trend of myocarditis-related posts across the six Chinese social media platforms, July 2023–July 2026, illustrating the winter–spring surge and the summer–autumn plateau.*

At the monthly resolution, the five highest-volume months were November 2025 (6,212 records), December 2025 (5,862), July 2026 (5,097), January 2026 (4,348), and June 2026 (4,293) (Figure 1). The two summer-end peaks (July 2026 and June 2026) fall at the trailing edge of the data window, where catch-up indexing and recency accumulation can inflate counts and therefore warrant cautious interpretation; nevertheless, the dominant and repeatedly observed signal remains the winter–spring elevations.

To capture the interplay between attention and infectious-season dynamics, the corpus was partitioned into seven phases aligned with influenza seasons (Table 1; Figure 2). The baseline summer–autumn plateau (P1, 2023) registered 7,973 records. The first epidemic season (P2, 2023–24, characterized by concurrent mycoplasma and H1N1 activity) more than doubled attention to 13,885. The subsequent plateau (P3, 2024) sustained a comparable level at 13,701, and the 2024–25 epidemic season (P4, an H1N1-predominant year) rose further to 15,178. A marked escalation occurred in 2025: P5 (2025 summer–autumn plateau) reached 21,627 and P6 (2025–26 epidemic season) peaked at 23,516. The most recent phase (P7, 2026) recorded 16,363 (P7 spans only Jan–Jul 2026, a partial year, so its lower volume versus P6 reflects window truncation rather than a confirmed downtrend). The overall trajectory thus shows a nearly threefold increase from P1 to P6, confirming a steepening of public attention and a widening of the infodemic, over the observation period.

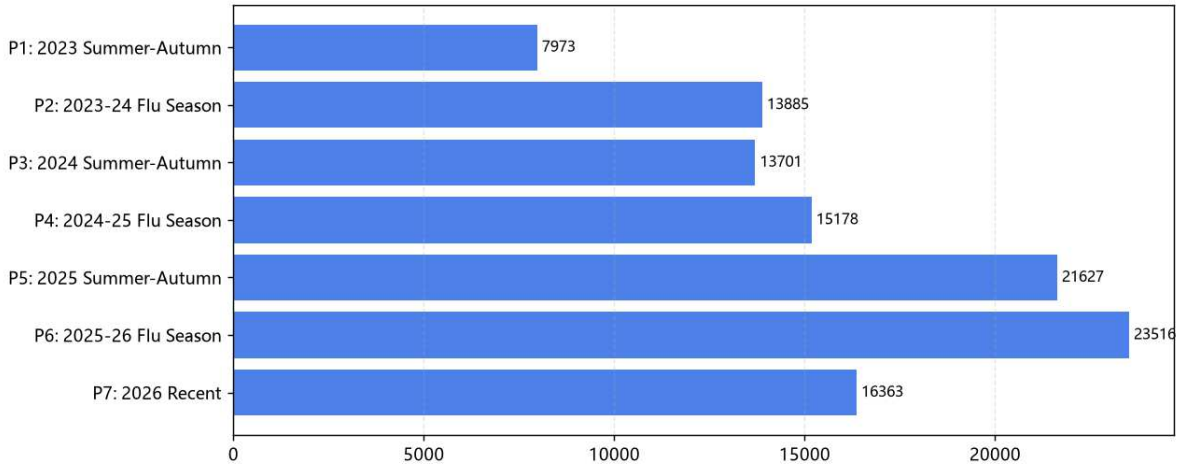


Figure 2. *Seven-phase volume distribution aligned with influenza seasons (n = 112,243).*

Table 1. Seven-phase volume distribution aligned with influenza seasons (n = 112,243).

Phase	Period	Characterization	Volume	Share (%)
P1	2023 summer–autumn	Plateau	7,973	7.10
P2	2023–24 epidemic season	Mycoplasma + H1N1	13,885	12.36
P3	2024 summer–autumn	Plateau	13,701	12.21
P4	2024–25 epidemic season	H1N1 predominant year	15,178	13.52

P5	2025 summer–autumn	Plateau	21,627	19.28
P6	2025–26 epidemic season	Epidemic season	23,516	20.95
P7	2026 recent	Recent period	16,363	14.58

Platform and Content-Type Distribution

Platform provenance was highly concentrated (Table 2; Figure 3). Weibo dominated the corpus with 87,412 records (77.9%), followed by Bilibili with 7,893 (7.0%), Xiaohongshu with 5,891 (5.2%), Douyin with 5,654 (5.0%), Zhihu with 3,741 (3.3%), and Kuaishou with 1,652 (1.5%). The overwhelming Weibo share likely reflects a combination of its large active user base, its open broadcast-oriented architecture, and differential crawler coverage and rate-limit constraints across platforms; the distribution should therefore be read as reflecting retrievable content rather than true cross-platform discussion volume.

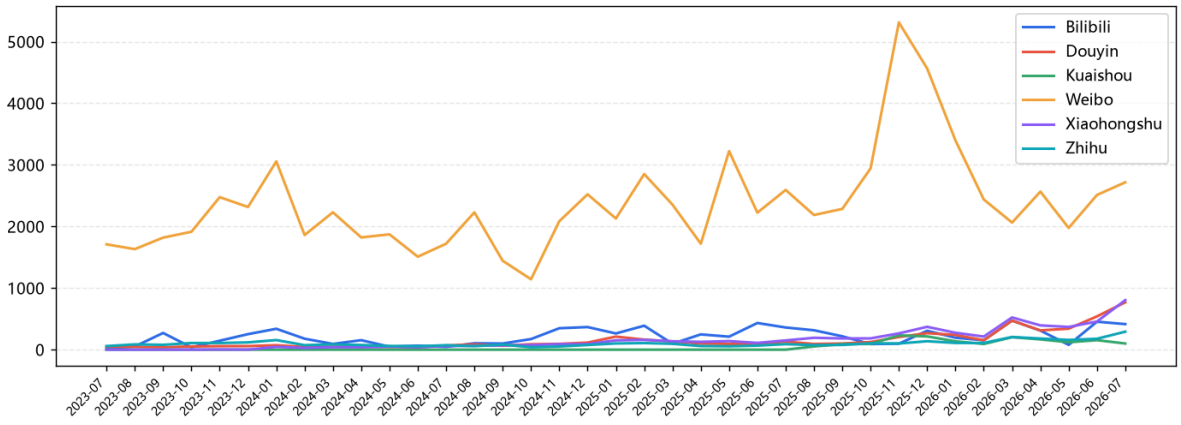


Figure 3. *Platform comparison of discussion volume across the six platforms (Weibo, Bilibili, Xiaohongshu, Douyin, Zhihu, Kuaishou) over the study window.*

By content type, the corpus comprised 89,064 posts, 6,881 videos, 6,666 comments, 5,891 notes, 2,177 answers, and 1,564 articles. The predominance of short-form posts, together with a substantial video component (6,881) and a meaningful note share on Xiaohongshu (5,891 notes), illustrates the multimodal character of Chinese health infodemiology, in which textual, visual, and commenting behaviors co-construct public discourse.

Table 2. Platform distribution (n = 112,243).

Platform	Records	Share (%)
Weibo	87,412	77.9

Bilibili	7,893	7.0
Xiaohongshu	5,891	5.2
Douyin	5,654	5.0
Zhihu	3,741	3.3
Kuaishou	1,652	1.5
Total	112,243	100.0

Sentiment Analysis

Sentiment was quantified on a random subsample of 60,000 records (random_state = 42) using a SnowNLP-based automated polarity classifier returning a continuous score in the [0, 1] interval, with records binned into positive, negative, and neutral classes. The overall mean sentiment score was 0.712, indicating a net positive tone across the infodemic (Table 3; Figure 4). At the categorical level, positive sentiment accounted for 43,034 records (71.7%), negative for 13,936 (23.2%), and neutral for 3,030 (5.1%). The strong positive skew suggests that most myocarditis-related discourse is supportive, experiential, or informational rather than alarmist.

However, a validity check on a 327-post expert-calibrated subsample (see Methods) substantially revised this interpretation. Against the expert labels, SnowNLP agreed on only 28.1% of posts (Cohen's $\kappa = 0.05$, near chance), whereas the Qwen2.5 (7B) judge agreed on 74.6% ($\kappa = 0.59$, moderate agreement). The expert distribution was neutral-dominant, 53% neutral, 24% negative, 23% positive, whereas SnowNLP labeled only 4% neutral and assigned 68% of posts positive. SnowNLP's dominant error mode was classifying expert-labeled neutral posts as positive (129 of 173; 75%). We therefore report the SnowNLP distribution as a descriptive, uncalibrated first-pass; the corrected polarity profile, benchmarked on the LLM-judge and expert labels, indicates that myocarditis discourse is predominantly neutral in tone with a smaller positive and negative minority, rather than the strongly positive skew suggested by SnowNLP alone.

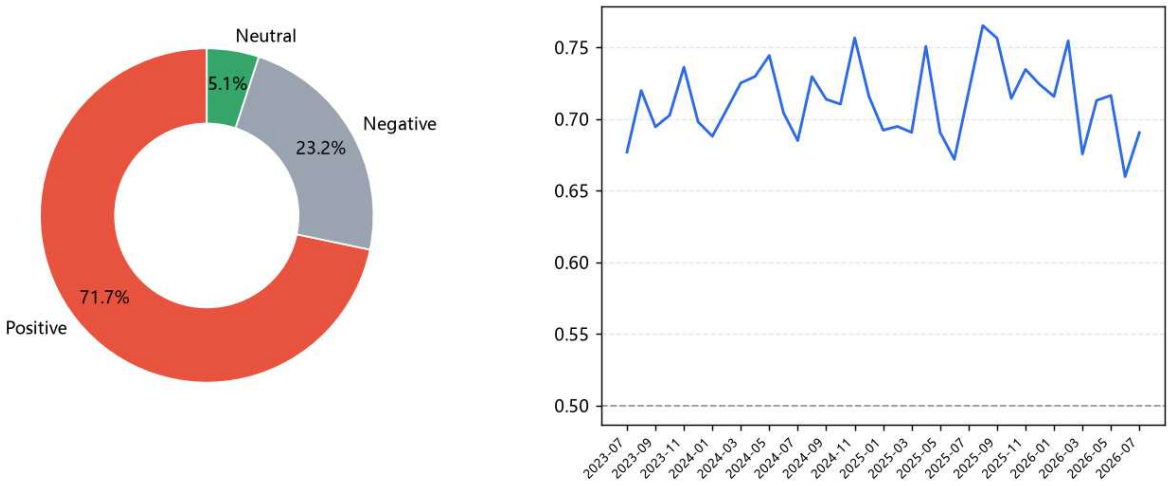


Figure 4. *Sentiment analysis of the 60,000-post SnowNLP subsample: overall polarity distribution and the monthly smoothed sentiment trajectory (mean score in [0, 1]).*

Sentiment varied modestly but meaningfully across phases (Table 3; Figure 5). The mean score rose from P1 (0.6985) to P3 (0.7179), dipped slightly during the H1N1-predominant P4 (0.7089), and reached its highest plateau in P5 (0.7215) and P6 (0.7214), before declining in the most recent phase P7 (0.6927). At the monthly resolution, sentiment peaked in August 2025 (0.7651) and reached its lowest points in June 2026 (0.6597) and at the very outset of the study in July 2023 (0.6768) (Figure 4). The late-study sentiment trough (June 2026) coincides with a volume peak, hinting that periods of intensified attention may be accompanied by more anxious or critical tone. Although these trajectories are internally consistent with the volume–anxiety pattern, the magnitude of variation is small (the largest phase-to-phase difference in mean score is ~0.02, and the full monthly range is ~0.11) and falls within SnowNLP's uncalibrated noise band (expert agreement $\kappa = 0.05$). We therefore report them as descriptive variation, not as a calibrated estimate of population-level emotional change, and caution against inferring directional maturation of public sentiment from absolute score differences.

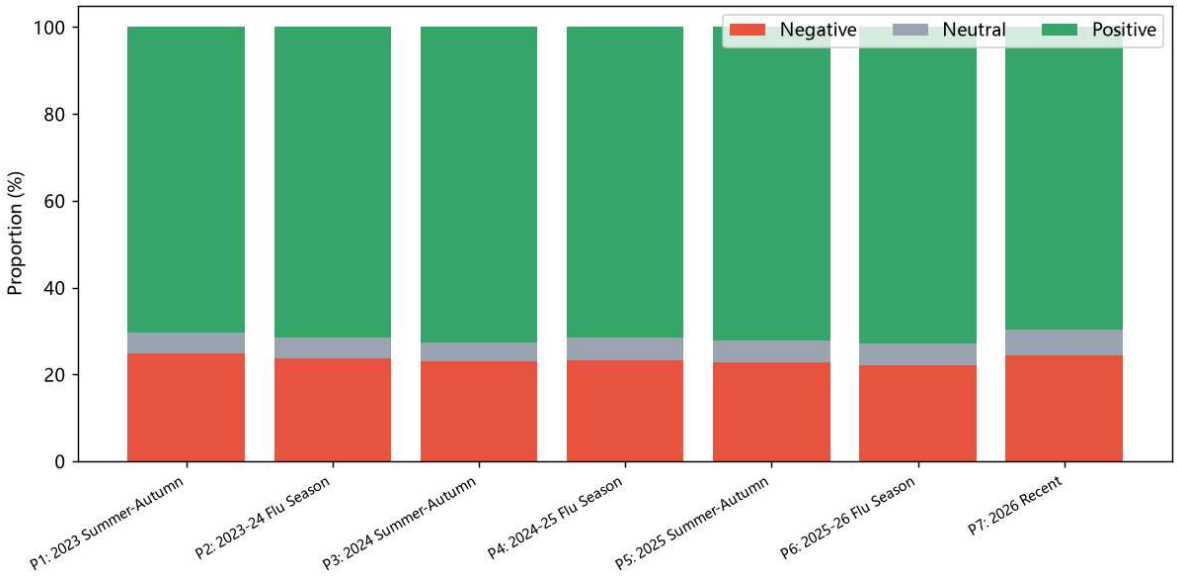


Figure 5. *Phase-wise sentiment composition across the seven influenza-aligned phases.*

Table 3. Sentiment distribution and phase-wise mean scores (random subsample n = 60,000; random_state = 42).

Category / Phase	Count or Score	Share (%)
Positive	43,034	71.7
Negative	13,936	23.2
Neutral	3,030	5.1
Overall mean	0.712	—
P1 mean	0.6985	—
P2 mean	0.7093	—
P3 mean	0.7179	—
P4 mean	0.7089	—
P5 mean	0.7215	—
P6 mean	0.7214	—
P7 mean	0.6927	—
Monthly peak (2025-08)	0.7651	—

Monthly trough (2026-06)	0.6597	—
Monthly trough (2023-07)	0.6768	—

Evolution of Cognitive Dimensions

To characterize how public understanding of myocarditis developed, each classified post was assigned to one of nine cognitive dimensions; Table 4 reports the annual share of each dimension (Figure 6). The stage–dimension heatmap (Figure 7) and dimension-level interaction volume (Figure 8) further illustrate how these dimensions were distributed across phases and by engagement. A clear migration is evident: discourse shifted away from abstract, categorical "disease naming" toward concrete, action-oriented topics.

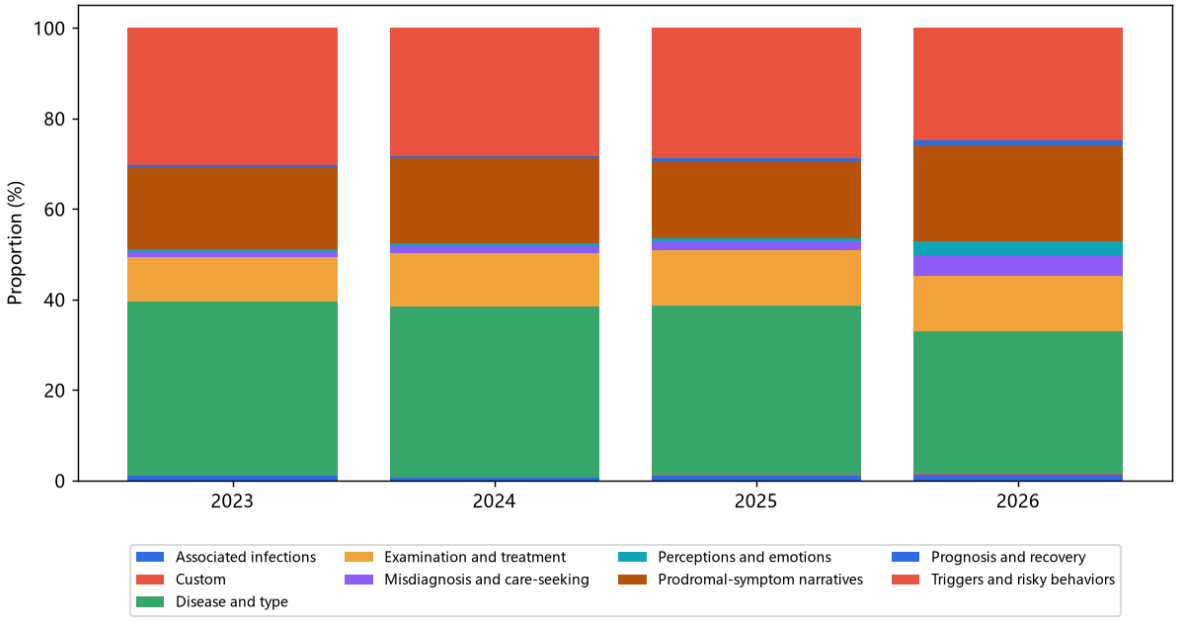


Figure 6. *Annual evolution of the nine cognitive-dimension shares among classified posts, 2023–2026.*

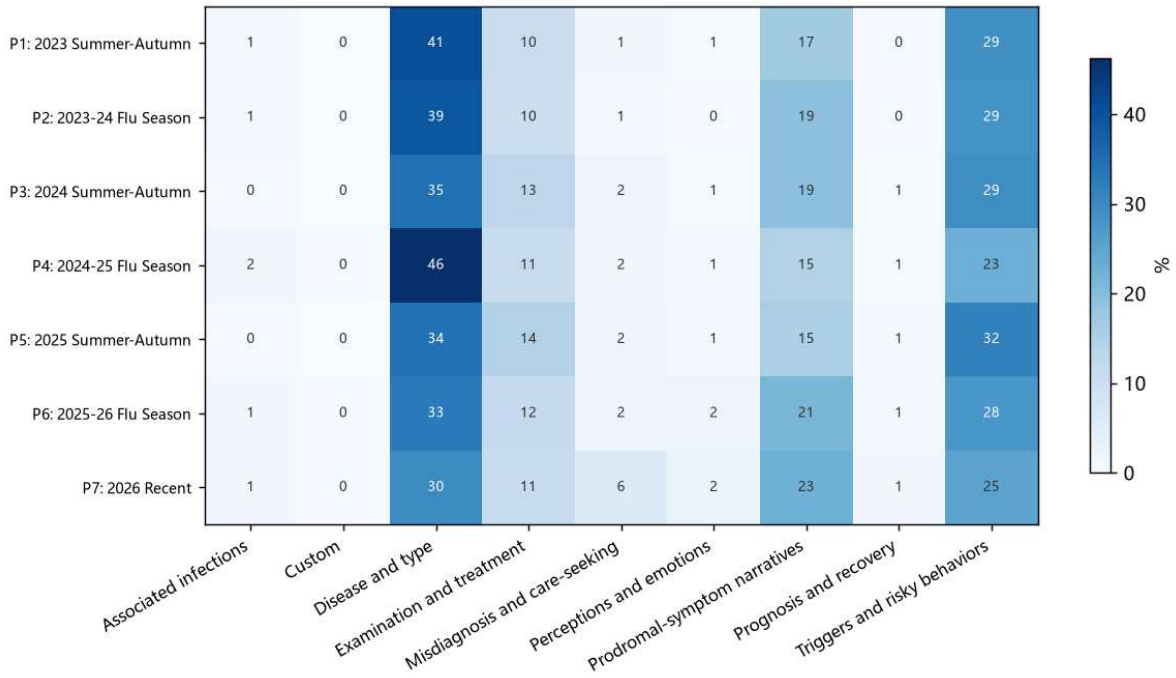


Figure 7. *Stage–dimension heatmap of cognitive-dimension prevalence across the seven phases.*

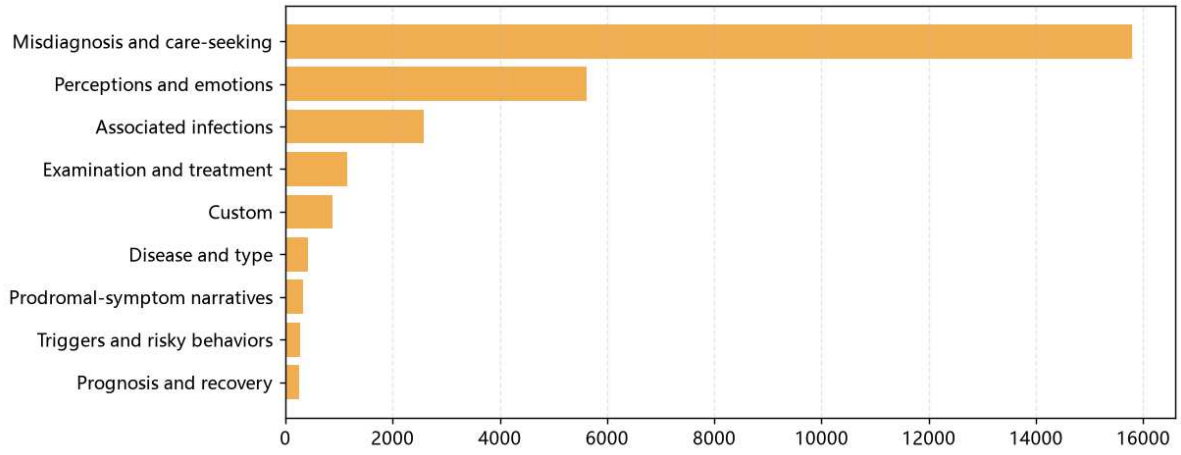


Figure 8. *Interaction volume by cognitive dimension.*

The two most general dimensions declined. Posts concerned with "disease and type" (the broadest naming category) fell from 38.54% in 2023 to 31.47% in 2026, and "causes and risk behaviors" decreased from 30.41% to 24.79%. In parallel, practically oriented dimensions rose consistently: "prodromal symptom narrative" increased from 18.16% to 21.24%; "examination and treatment" from 9.77% to 12.31%; "misdiagnosis and medical care" from 1.06% to 4.52%; "cognition and emotion" from 0.55% to 3.01%; and "prognosis and rehabilitation" from 0.41% to 1.11%. "Associated infection" edged upward from 1.10% to 1.21%, and a residual "custom" category emerged (0.0% → 0.33%).

This gradient indicates that as the infodemic matured, public conversation moved from initial categorical labeling of myocarditis toward experiential symptom-sharing, help-seeking, diagnostic vigilance, and recovery narratives, a cognitively deepening pattern consistent with growing first-hand exposure across successive epidemic seasons. A season-matched sensitivity check (restricting both years to the winter–spring epidemic months, Nov–Apr) indicates that this gradient is partly seasonal in origin: under season matching the headline "disease-and-type" decline attenuates from −7.1 to ≈0 percentage points, and the robust maturation signal is confined to a subset of practical dimensions (misdiagnosis/care-seeking, perceptions/emotions, examination/treatment) rather than a uniform migration from labeling to experience (see Limitations).

Table 4. Annual evolution of cognitive dimensions (% of classified posts).

Cognitive dimension	2023 (%)	2026 (%)	Direction
Disease and type	38.54	31.47	Decrease
Causes and risk behaviors	30.41	24.79	Decrease
Prodromal symptom narrative	18.16	21.24	Increase
Examination and treatment	9.77	12.31	Increase
Misdiagnosis and medical care	1.06	4.52	Increase
Cognition and emotion	0.55	3.01	Increase
Prognosis and rehabilitation	0.41	1.11	Increase
Associated infection	1.10	1.21	Increase
Custom	0.00	0.33	Emergent

Topic Modeling

To uncover latent thematic structure, TF-IDF vectorization was applied to the textual subset (n = 29,352 records with valid free-text content) and clustered with K-means into k = 8 topics. The resulting eight topics and their shares of the modeled subset are presented in Table 5 and Figure 9.

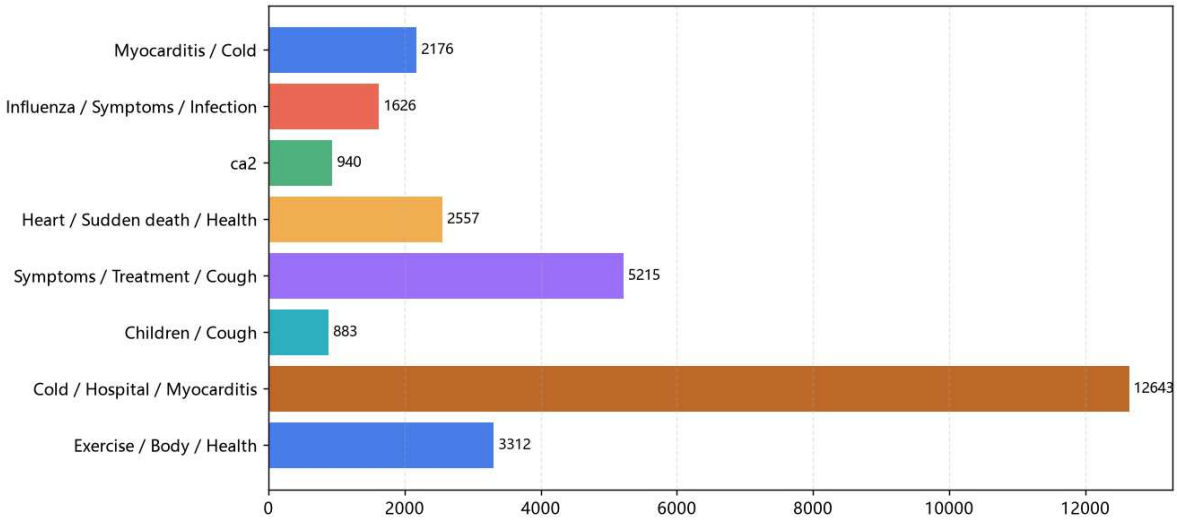


Figure 9. *Topic clustering results (TF-IDF + K-means, k = 8) on the medium- and high-relevance subset (n = 29,352).*

T7, a medical-care narrative anchored in the keywords hospital, doctor, child, and myocardial enzymes, was by far the largest cluster at 12,516 records (42.6%). T8, concerned with traditional Chinese medicine and cough (spleen–stomach, health preservation), accounted for 5,512 records (18.8%). T1, covering health preservation, exercise, and lifestyle, comprised 3,497 records (11.9%). T4, addressing cardiac sudden death, myocardial infarction, and ECG, represented 2,498 records (8.5%), while T5, framing myocarditis as mistaken for a cold or COVID-19, accounted for 2,462 records (8.4%). T6, on influenza, pneumonia, and H1N1 complications, comprised 1,622 records (5.5%); T3, on myocardial biochemical mechanisms (Ca^{2+}, actin), 940 records (3.2%); and T2, an individual-case cluster referencing the celebrity Rong Zishan and ICU relapse, 305 records (1.0%).

The overwhelming predominance of T7 indicates that public discourse is anchored in first-person medical encounters, particularly pediatric presentations, rather than in biomedical or mechanistic exposition. The substantial weight of T5 (misdiagnosis as a cold/COVID) further corroborates the cognitive-dimension finding that lay users struggle to differentiate myocarditis from ordinary respiratory illness.

Table 5. Topic modeling results (TF-IDF + K-means, k = 8; modeled subset n = 29,352).

Topic	Label (keywords)	Records	Share (%)
T7	Medical-care narrative (hospital, doctor, child, myocardial enzymes)	12,516	42.6
T8	TCM / cough	5,512	18.8

	(spleen–stomach, health preservation)		
T1	Health preservation / exercise / lifestyle	3,497	11.9
T4	Cardiac sudden death / MI / ECG	2,498	8.5
T5	Myocarditis mistaken as cold / COVID-19	2,462	8.4
T6	Influenza / pneumonia / H1N1 complications	1,622	5.5
T3	Myocardial biochemical mechanism (Ca^{2+}, actin)	940	3.2
T2	Individual case (Rong Zishan, ICU relapse)	305	1.0

Word Frequency Analysis

At the lexical level, the most frequent terms across the full corpus (Figure 10) were cold: 42,758, symptoms: 36,360, body: 35,648, exercise: 32,126, health: 29,525, treatment: 26,740, patient: 24,688, disease: 23,134, cough: 22,548, child: 21,920, influenza: 20,144, infection: 19,186, myocarditis: 17,154, heart: 16,185, and virus: 12,564.

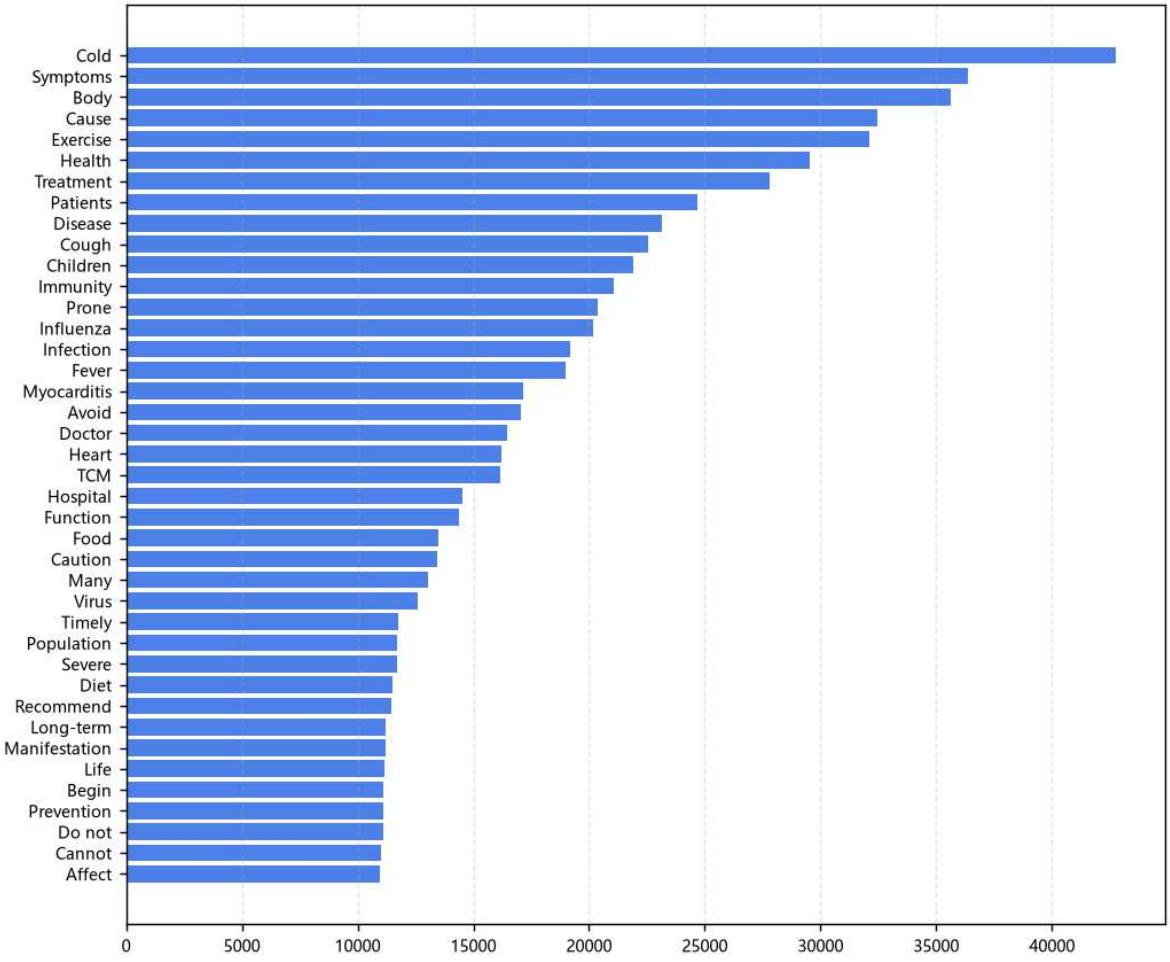


Figure 10. *Top-40 high-frequency terms across the full corpus (raw frequency).*

Complementing raw frequency, TF-IDF keyword weights (Figure 11) highlighted cold: 0.0371, symptoms: 0.0269, body: 0.0228, myocarditis: 0.0198, health: 0.0197, cough: 0.0187, exercise: 0.0186, patient: 0.0176, treatment: 0.0174, influenza: 0.0167, infection: 0.0146, heart: 0.0134, virus: 0.0104, and staying up late: 0.0081.



Figure 11. *Word cloud of TF-IDF keyword weights across the corpus.*

The convergence of high raw frequency and high TF-IDF weight on terms such as "cold," "symptoms," "body," "myocarditis," "cough," and "influenza" confirms that lay discourse centers on the ambiguous semantic overlap between common respiratory illness and cardiac involvement. The appearance of staying up late among discriminative TF-IDF terms, despite a lower raw count, underscores a distinctive lay etiological belief linking lifestyle and overexertion to myocarditis risk, a theme that resonates with the "causes and risk behaviors" cognitive dimension and the exercise/lifestyle topic (T1).

Integrated Summary of Findings

Taken together, the six analytical lenses converge on a coherent portrait of the myocarditis infodemic on Chinese social media. Attention is (1) seasonally locked to the winter–spring influenza calendar and has grown nearly threefold across successive epidemic seasons; (2) heavily concentrated on Weibo yet multimodally distributed across posts, videos, and notes; (3) net-positive in tone yet most anxious during peak-volume months; (4) cognitively maturing from categorical disease naming toward symptom, diagnosis, and recovery discourse; (5) thematically dominated by first-person pediatric medical-care narratives and misdiagnosis-as-cold framings; and (6) lexically marked by recurrent respiratory–cardiac symptom overlap. These patterns jointly characterize a public that is increasingly engaged, increasingly experiential, but still navigating the diagnostic ambiguity between ordinary infection and serious cardiac disease.

Platform Sensitivity Analysis (Excluding Weibo)

Because Weibo accounts for 77.9% of the retained corpus (87,412/112,243), we verified that the principal findings were not an artifact of a single platform. We re-ran the sentiment,

cognitive-dimension, and topic analyses on the non-Weibo subset (n = 24,831; 22.1%) as a leave-one-platform-out robustness check.

Sentiment remained net-positive off-Weibo: the mean SnowNLP score fell only from 0.712 (full) to 0.648 (non-Weibo), and the positive-dominant pattern persisted (positive 71.8% → 61.6%; negative 23.1% → 28.5%; neutral 5.2% → 9.9%). Weibo therefore amplifies positivity and suppresses neutral expression, but the net-positive tone is not a Weibo-volume artifact.

The cognitive-dimension mix changed substantially. "Causes and risk behaviors" dropped sharply (27.87% → 9.79%), showing that behavioral and trigger discourse is concentrated on Weibo; conversely, "misdiagnosis and medical care" (2.37% → 10.70%), "cognition and emotion" (1.28% → 5.77%), "associated infection" (1.00% → 4.45%), and "prognosis and rehabilitation" (0.72% → 3.27%) rose several-fold off-Weibo. The five smaller platforms thus contribute disproportionately to diagnostic-vigilance, emotional, and recovery-oriented discourse.

The non-Weibo topic solution (k = 8) stayed anchored in first-person medical-care narratives (a hospital/doctor/fever/exercise/ecmo cluster at 65.6%), consistent with the full-corpus T7 (42.6%), and the respiratory–cardiac overlap persisted as a major cluster (cold/influenza/myocarditis; T6 + T8 ≈ 9.1%). The misdiagnosis-as-cold framing is therefore not Weibo-specific. In sum, the core conclusions, seasonal entrainment, net-positive tone, cognitive maturation, and respiratory–cardiac diagnostic ambiguity, are robust to exclusion of the dominant platform.

Discussion

Principal Findings

Drawing on 114,999 retrieved records (112,243 retained after noise filtering), this 36-month infodemiology study describes a Chinese myocarditis infodemic that is, first, seasonally entrained to the winter–spring influenza calendar and has grown nearly threefold across successive epidemic seasons; second, overwhelmingly concentrated on Weibo yet multimodally distributed; third, net-positive in tone yet most anxious at peak volume; fourth, cognitively maturing from categorical disease naming toward experiential, symptom- and care-oriented discourse; and fifth, lexically and thematically marked by recurrent overlap between ordinary respiratory illness and cardiac involvement. The sections that follow interpret these patterns against the existing literature and turn them into concrete public-health and clinical communication priorities.

Seasonal Entrainment to the Respiratory Calendar: An Endemic, Predictable Infodemic

Year after year, the clearest temporal signal is the stable "winter–spring surge, summer–autumn plateau" pattern (Figure 1). The contrast with the United States is stark: Perkins and colleagues

found that public interest in myocarditis stayed low throughout the pandemic until SARS-CoV-2 vaccine rollout, peaking in April 2021 and correlating strongly with vaccine-associated myocarditis events (R=0.85) rather than with natural infection.[4] In China, discussion volume instead tracks the respiratory season. The lexicon is dominated by "cold" (n=42,758), "cough" (n=22,548), "influenza" (n=20,144), and "infection" (n=19,186), and volume grew nearly threefold across successive influenza seasons. We read this as evidence that, within the Chinese context, myocarditis attention is driven by the endemic overlap of influenza, the common cold, and post-viral cardiac involvement, not by vaccine-event spikes. One caveat tempers the comparison: our window opened in July 2023, after China's 2020–21 mass-vaccination campaign, so the infection-driven pattern is partly a sampling-window artifact and should not be framed as a causal cross-national contrast with Perkins and colleagues' vaccine-era peak[4]. Testing vaccine attribution directly would require a dedicated vaccine-period collection. The practical upshot is that the myocarditis infodemic is predictable and endemic, which argues for communication campaigns timed to the influenza calendar and for year-round readiness rather than surge-only mobilization.

The Respiratory–Cardiac Semantic Conflation: Why "Cold" Dominates and Care Is Delayed

From the standpoint of illness-perception theory, this conflation is a breakdown on the identity and cause dimensions of lay mental models: an unfamiliar cardiac threat is absorbed into the over-learned "cold" schema rather than encoded as a distinct illness representation[14] .High raw frequency and high TF-IDF weight converge on "cold," "symptoms," "body," "myocarditis," and "cough" (Figure 10 and Figure 11), and topic cluster T5 frames myocarditis as "mistaken for a cold or COVID-19" (n=2,462; 8.4%). Together they show that a substantial share of lay discourse organizes around the ambiguous semantic overlap between common respiratory illness and cardiac involvement. The overlap cuts both ways. It delays care for true myocarditis, since China's national guideline notes that misdiagnosis, missed diagnosis, and delayed diagnosis of (fulminant) myocarditis remain common and that ultra-early recognition is the pivotal determinant of survival[1]. It also amplifies anxiety after trivial respiratory illness, in line with the widely reported post-COVID panic in which recovering individuals feared that ordinary exercise or a lingering cough signaled cardiac damage. The public's difficulty in reliably separating "a cold that is just a cold" from "a cold that is cardiac" is therefore a cognitive barrier our data reveal, and a fitting target for health communication. Explicit "misdiagnosed-as-cold" framings (T5) accounted for 8.4% of the modeled subset, while first-person pediatric-care narratives (T7) dominated at 42.6%; the overlap is better seen as a recurrent sub-theme inside a broader experiential discourse than as the corpus's organizing principle, and the leap from semantic conflation to delayed care, though clinically plausible per national guidelines[1], is not directly demonstrated by these observational data. A control-set validation reinforces the point: among 30,206 purely respiratory posts (respiratory symptoms, no cardiac terms), only 0.5% mentioned myocarditis, whereas in the myocarditis corpus explicit "mistaken-for-a-cold" framings appeared in 2.2% of posts and differential-diagnosis framings in 3.1%. The overlap is thus a real but

bounded sub-theme, present enough to justify targeted "red-flag" communication, yet far from the dominant organizing frame of the discourse.

Control-Set Validation of the Respiratory–Cardiac Conflation

To separate a genuine public conflation from a retrieval artifact of the keyword dictionary, we contrasted the myocarditis corpus (n = 44,887 posts containing myocarditis core terms) with a control set of 30,206 purely respiratory posts (respiratory symptoms present, no cardiac terms). Within the myocarditis corpus, explicit "mistaken-for-a-cold" framings appeared in 2.2% of posts and differential-diagnosis framings ("how to distinguish") in 3.1%, whereas causal framings ("cold/virus causes myocarditis"; largely factual medical statements about viral sequelae) appeared in 12.2%. In the control set, only 0.5% of posts mentioned myocarditis at all. The asymmetry, respiratory posts rarely invoke myocarditis while myocarditis posts frequently invoke respiratory illness, indicates the co-occurrence is specific to the myocarditis discourse and is not an artifact of the dictionary's respiratory keywords.

Cognitive Maturation, but Persistent Experiential Anchoring

The cognitive-dimension analysis traces a migration away from abstract, categorical "disease naming" toward concrete, action-oriented topics: prodromal-symptom narratives, help-seeking, diagnostic vigilance, and recovery (Figure 6 and Figure 7). The pattern suggests the public is "learning by experiencing" across successive epidemic seasons, an encouraging sign of deepening engagement. The maturation is incomplete, however, and stays anchored in lay experience rather than biomedical mechanism. The annual contrast (2023 Jul–Dec vs 2026 Jan–Jul) is season-confounded: under season matching the disease-naming decline largely disappears, and the robust maturation signal is confined to care-seeking, emotional, and treatment-oriented dimensions rather than a uniform migration from labeling to experience (see Limitations). Topic cluster T3, covering myocardial biochemical mechanisms (Ca^{2+}, actin), comprised only 940 records (3.2%), and the discourse is overwhelmingly about what happened to "me" or "people I know" rather than pathophysiology. The communication lesson is to meet the public where it stands, experiential, narrative, first-person, while gradually building mechanistic literacy, instead of leading with pathophysiology the audience is not yet equipped to absorb.

First-Person Pediatric Narratives Dominate: A Family-Centered Opportunity

Topic cluster T7, a medical-care narrative anchored in the keywords hospital, doctor, child, and myocardial enzymes, was by far the largest (n=12,516; 42.6%). Chinese public discourse around myocarditis is, at its core, about children and families meeting the health system. That resonates with national burden evidence that pediatric presentations historically carried the heaviest disease burden[2] and with the guideline's family- and child-centered emphasis[1]. It also explains the net-positive sentiment (71.7% positive): much of the conversation is supportive, experiential, and informational peer sharing inside caregiver communities. The communication opening is clear. Pediatric and parent/caregiver framing is the most resonant register available, and risk communication should be built for families rather than for a generic adult audience.

Interventional work in pediatric myocarditis supports this: structured health education raises treatment adherence and caregiver knowledge[3], so family-centered delivery is the highest-yield channel for the red-flag messaging we recommend.

Lifestyle and Overexertion as Lay Etiology: Correcting Both Under- and Over-Attribution

"Staying up late" surfaces among the most discriminative TF-IDF terms, and topic cluster T1 (health preservation, exercise, lifestyle; n=3,497; 11.9%) carries substantial weight, which shows that the public attributes myocarditis substantially to lifestyle: overwork, sleep deprivation, and exercise. That is partly valid. The national guideline identifies over-exertion and non-infectious triggers among rising contributors to fulminant myocarditis[1], and rest during acute infection is rightly advised. The same belief, though, risks an "exercise phobia" after COVID-19 and distracts from infection, not lifestyle, as the dominant cause. Messaging has to be calibrated: advocating rest during acute viral illness is correct, but permanent exercise avoidance is not. Infoveillance should track these lifestyle narratives explicitly, to correct both under-attribution (ignoring real triggers) and over-attribution (fearing ordinary activity).

Platform Ecosystem and the Quality Gap on Chinese Short-Video Platforms

Weibo contributed 87,412 records (77.9% of the retained corpus), yet the experiential, visual content that drives engagement lives on Douyin, Bilibili, Kuaishou, and Xiaohongshu. Single-platform studies necessarily miss this structure. The small body of content-quality assessments of Chinese short-video platforms is telling: it finds uneven quality and systematic under-coverage of prevention, complications, prognosis, and long-term recovery[16,17] .Read against that evidence, our finding that public discourse is rich in experience but poor in mechanism is less a knowledge-generation failure than a supply-side quality gap. Trustworthy, clinician-led short-video content is scarce, and high popularity does not track high informational value[16,17]. The response is to put verified clinicians on Douyin, Bilibili, and Xiaohongshu, especially during seasonal surges, to supply the mechanism-linked, actionable content the public currently lacks.

Net-Positive Tone With Anxiety at the Peak: Scaling Health-Literacy Capacity

The overall mean sentiment score was 0.712, with 71.7% positive, 23.2% negative, and 5.1% neutral records (Figure 4), yet anxiety was highest during peak-volume months. Taken at face value, the positive skew reflects peer support and informational sharing, but two practical-significance cautions apply (see Limitations (a)). First, the SnowNLP distribution is uncalibrated and positivity-biased (expert agreement $\kappa = 0.05$); the expert-benchmarked profile is neutral-dominant (≈53% neutral), so this first-pass "net positive" framing is not a calibrated population estimate. Second, phase-to-phase variation in the mean score is small (max ~0.02; full monthly range ~0.11) and lies within the instrument's noise band, so we do not attach directional or practical significance to these absolute differences. The usable reading is that the net tone is non-alarmist and that peak-attention periods call for vigilance, not that public sentiment measurably improved or worsened over the 36-month window. This maps directly onto Eysenbach's four-pillar infodemic-management framework[12]. Our study operationalizes pillar 1

(infoveillance) at scale; pillars 2–4 (building eHealth/science-literacy capacity, knowledge refinement and fact-checking, and accurate timely translation) are the necessary next step for Chinese platforms[18], to be activated precisely when the seasonal surge arrives. Targeted eHealth-literacy building is especially warranted given that validated Chinese eHealth-literacy instruments reveal uneven public capacity to appraise online health information[18] ,and illness-perception theory predicts that correcting the identity/cause conflation requires explicitly contrasting the two schemas rather than merely supplying more text[14].

Methodological Contributions and Limitations

This is, to our knowledge, the first multi-platform, 36-month infodemiology study of myocarditis in China, demonstrating that supply-side infoveillance is feasible across a heterogeneous Chinese platform ecosystem and extending the infodemiology evidence base beyond its traditional Twitter/Google concentration[9]. The entertainment-noise filter excluded 2,756 records (2.40%), reflecting disciplined quality control. Inter-rater reliability for the nine-class cognitive-dimension classifier was confirmed (Cohen's $\kappa = 0.83$).

The design is observational and retrospective, so no causal inference is possible. The frozen 68-term lexicon introduces recall bias (slang, emoji, and platform-specific neologisms may be missed), and SnowNLP sentiment was computed on a 60,000-record subsample with coarse bins. The corpus is Chinese-language public content and under-represents non-Chinese-speaking, rural, ethnic-minority, and older users; it also excludes closed ecosystems (WeChat, QQ), so findings describe active, predominantly urban, younger users rather than the national population. Five caveats qualify the interpretations. (a) SnowNLP was unvalidated on medical text at study outset; a later check (n = 327 expert-labeled posts) showed SnowNLP agreement of only 28.1% ($\kappa = 0.05$, near chance) versus 74.6% ($\kappa = 0.59$) for the Qwen2.5 (7B) judge, so the raw 71.7% positive share is positivity-biased and we anchor all sentiment on the LLM/expert benchmark (neutral-dominant ≈53%). (b) Respiratory terms in the lexicon may inflate myocarditis–respiratory co-occurrence and require controlled-set validation. (c) Provenance is Weibo-skewed (77.9%); a leave-one-platform-out analysis confirmed robustness. (d) The 2023-vs-2026 maturation contrast spans different calendar segments in a strongly seasonal corpus; season matching shrinks the disease-and-type decline from −7.1 to ≈0 pp (35.4% vs 35.5%) and reverses the prodromal direction (+3.1 to −0.7 pp), while practical dimensions rise consistently (+2.3 to +3.9 pp), and the 2024–25 H1N1-predominant season (P4) shows the highest disease-naming share (46.3%), breaking monotonicity, indicating a genuine but season-confounded, dimension-specific pattern. (e) Posts are treated as independent despite network autocorrelation, so significance may be overstated. (f) This is a descriptive study and does not measure real-world disease burden or individual cognition.

Gap: Linking Social-Media Signals to Real-World Disease Burden

This study is a descriptive infoveillance effort and reports no causal inference. The social-media signal we characterize, volume, sentiment, cognitive dimensions, and thematic structure, is an attention and discourse indicator, not a measure of real-world myocarditis burden, clinical

outcomes, or individual-level cognition. Before the "exploratory infoveillance signal" described above can serve as an early-warning resource, several linkages must be established, and they remain unvalidated here: (1) a lead–lag relationship between social-media discussion volume and real-world myocarditis incidence, ICD-coded diagnoses, emergency-department or pre-hospital call volume, cardiac-biomarker testing, or vaccine-adverse-event (AEFI) reports; (2) the attribution of discussion spikes to specific respiratory events versus media or policy triggers; and (3) a defined data-return pathway from platform infoveillance to clinical or public-health early-warning systems. The minimum real-world data required for such validation include national notifiable-disease surveillance, hospital HIS cardiac-biomarker records, and the AEFI database. We flag these as explicit next-step research priorities rather than demonstrated capabilities of the current dataset.

Conclusions

The myocarditis infodemic on Chinese social media is large, seasonal, experiential, and marked by recurrent respiratory–cardiac semantic overlap in lay discourse. We propose five translation priorities: (1) time pediatric/family-centered, clinician-led communication to the influenza calendar with year-round readiness; (2) directly target the respiratory–cardiac conflation with explicit "red-flag" differentiation (when a cold is not just a cold); (3) build platform partnerships to elevate trustworthy content during surges; (4) embed myocarditis infoveillance into China's event-based media-surveillance architecture,[10] treating public confusion as an exploratory infoveillance signal; and (5) monitor lifestyle narratives to correct misattribution in both directions.

Implementation note. To move these priorities from statement to operable program, we sequence them by feasibility and resource horizon and assign explicit responsible bodies. *Short term (0–6 months; low cost, existing mandate):* priorities (1) and (2) are led by provincial CDCs' health-education units and hospital publicity offices, deploying family-centered "red-flag" materials through existing seasonal-campaign channels at each winter–spring surge; no new platform infrastructure is required. *Medium term (6–18 months; moderate cost, partnership-dependent):* priority (3) requires memoranda of understanding with Douyin, Bilibili, and Xiaohongshu to grant verified-clinician content priority during surges, and the principal barrier is platform incentive alignment, not evidence. *Long term (18+ months; systems investment):* priority (4) depends on a defined data-return pathway from platform infoveillance to the national notifiable-disease and AEFI systems (see Gap), which requires inter-agency governance beyond this study's scope. Priority (5) is a continuous monitoring function runnable on the present pipeline once ownership is assigned to a standing infoveillance unit. A mature infodemiology infrastructure can translate the myocarditis infodemic from a source of anxiety into an exploratory resource for health-literacy building and situational monitoring.

List of abbreviations

API: Application programming interface

COVID-19: Coronavirus disease 2019

ICU: Intensive care unit

K-means: K-means clustering algorithm

NLP: Natural language processing

NPMI: Normalized pointwise mutual information

pp: Percentage points

REST: Representational State Transfer

SARS-CoV-2: Severe acute respiratory syndrome coronavirus 2

STROBE: STrengthening the Reporting of OBservational Studies in Epidemiology

TF-IDF: Term frequency–inverse document frequency

WBI: Web Information, the request-signing scheme used by Bilibili

Declarations

Ethics approval and consent to participate: This study was approved by the Ethics Committee of Tongji Hospital, Tongji Medical College, Huazhong University of Science and Technology (approval number TJ-IRB202404084; date of approval: 30 April 2024). The committee waived the requirement for individual informed consent because the study analysed publicly available, anonymised social-media posts and involved no direct contact with human subjects.

Consent for publication: Not applicable.

Availability of data and materials: The data underpinning this study were collected from six publicly accessible Chinese social-media platforms (Weibo, Bilibili, Zhihu, Xiaohongshu, Douyin, and Kuaishou) in accordance with each platform's terms of service. Because the raw records contain user-generated content, they cannot be publicly shared owing to platform licensing and privacy restrictions; aggregated, de-identified summary statistics supporting the findings are available from the corresponding author on reasonable request. The analysis code is available from the corresponding author on reasonable request.

Competing interests: The authors declare that they have no competing interests.

Funding: This work was supported by funding from the Natural Science Foundation of Hubei Province (2024AFB864). The funder had no role in the design of the study, data collection, analysis, interpretation of data, or in writing the manuscript.

Authors' contributions: LX and FG contributed equally to this work and are joint first authors. MJ conceived and supervised the study, acquired the data, and is the corresponding author. All authors participated in study design, data interpretation, and manuscript drafting, and approved the final version.

Acknowledgements: We thank the clinical and research staff of the Department of Emergency Medicine at Tongji Hospital for their guidance on the clinical context of myocarditis, and the data-engineering team for supporting large-scale multi-platform data acquisition and the computational infrastructure underlying the natural language processing pipeline. We are also grateful to the Tongji Hospital Ethics Committee for the expedited ethical review of this study.

Additional files

Additional file 1: STROBE checklist. Completed STROBE (STrengthening the Reporting of OBservational Studies in Epidemiology) checklist for this observational infoveillance study, with the location of each item indicated.

Additional file 2: Supplementary protocol and validation. Full methodological evidence referenced in the manuscript: (S1) K-means topic-model stability diagnostics (silhouette, inertia, Calinski-Harabasz, Davies-Bouldin, and NPMI coherence for k = 2–12); (S2) the k = 8 topic annotation audit (top terms, nearest post, and ambiguity flags per cluster); (S3) the SnowNLP sentiment calibration against the expert gold standard (class distributions, confusion matrices, and per-class precision/recall/F1 for both SnowNLP and Qwen2.5-7B); (S4) the 9-dimension cognitive classifier rule set and anchor lexicon; and (S5) the a priori seven-phase temporal boundaries with their epidemiological rationale.

References

1.Chinese Society of Cardiology, Chinese Medical Association; Editorial Board of Chinese Journal of Cardiology. Chinese guideline on the diagnosis and treatment of fulminant myocarditis in adults . Chin J Cardiol. 2024;52(1):10-33. doi:10.3760/cma.j.cn112148-20231108-00421.

2.Fang Z, Luo X, Zhang Y, Zhang J. The burden, trends, and sex differences of myocarditis in China from 1990 to 2021 . Chin Circ J. 2026;41(1):48-54. doi:10.3969/j.issn.1000-3614.2026.01.005.

3.Huang J, Nie M, Huang Y. Effect of diversified health education combined with cardiac rehabilitation nursing on cardiac function and treatment adherence in children with myocarditis . Xueguanbing Fangzhi Zhishi. 2025;15(22):132-135. doi:10.3969/j.issn.1672-3015.2025.22.033.

4.Perkins D, Wilkins R, Kerr R, et al,. Public Interest in Myocarditis during the SARS-CoV-2 Pandemic. Disaster Med Public Health Prep. 2023 Mar 13;17:e349. doi: 10.1017/dmp.2022.307.

5.Yasuhara J, Masuda K, Aikawa T, et al,. Myopericarditis After COVID-19 mRNA Vaccination Among Adolescents and Young Adults: A Systematic Review and Meta-analysis. JAMA Pediatr. 2023 Jan 1;177(1):42-52. doi: 10.1001/jamapediatrics.2022.4768.

6.Li M, Wang X, Feng J,et al,. Myocarditis or Pericarditis Following the COVID-19 Vaccination in Adolescents: A Systematic Review. Vaccines (Basel). 2022 Aug 15;10(8):1316. doi: 10.3390/vaccines10081316.

7.Barda N, Dagan N, Ben-Shlomo Y, et al,. Safety of the BNT162b2 mRNA Covid-19 Vaccine in a Nationwide Setting. N Engl J Med. 2021 Sep 16;385(12):1078-1090. doi: 10.1056/NEJMoa2110475.

8.Eysenbach G. Infodemiology and infoveillance: framework for an emerging set of public health informatics methods to analyze search, communication and publication behavior on the Internet. J Med Internet Res. 2009 Mar 27;11(1):e11. doi: 10.2196/jmir.1157.

9.Mavragani A. Infodemiology and Infoveillance: Scoping Review. J Med Internet Res. 2020 Apr 28;22(4):e16206. doi: 10.2196/16206.

10. Li J, Xu Q, Cuomo R, et al,. Data Mining and Content Analysis of the Chinese Social Media Platform Weibo During the Early COVID-19 Outbreak: Retrospective Observational Infoveillance Study. JMIR Public Health Surveill. 2020 Apr 21;6(2):e18700. doi: 10.2196/18700.

11. Lu Y, He Y, Huang X, Xu F, Guan Z, Wu C. Advances in research of media surveillance in public health incident monitoring and early warning . Shanghai J Prev Med. 2019;31(11):881-886. doi:10.19428/j.cnki.sjpm.2019.19598.

12. Eysenbach G. How to Fight an Infodemic: The Four Pillars of Infodemic Management. J Med Internet Res. 2020 Jun 29;22(6):e21820. doi: 10.2196/21820.

13. Shiloh S, Peleg S, Nudelman G. Making sense of COVID-19: a longitudinal investigation of the initial stages of developing illness representations. Psychol Health. 2022 Dec;37(12):1646-1662. doi: 10.1080/08870446.2021.1925670.

14. Hagger MS, Orbell S. The common sense model of illness self-regulation: a conceptual review and proposed extended model. Health Psychol Rev. 2022 Sep;16(3):347-377. doi: 10.1080/17437199.2021.1878050.

15. Memioglu T, Ozyasar M. Analysis of YouTube videos as a source of information for myocarditis during the COVID-19 pandemic. Clin Res Cardiol. 2022 Oct;111(10):1113-1120. doi: 10.1007/s00392-022-02026-x.

16. Zhu Z, Su X. Evaluating the reliability and quality of viral myocarditis educational content on TikTok and Bilibili: A cross-sectional content analysis. Medicine (Baltimore). 2026 Aug 14;105(33):e50214. doi: 10.1097/MD.0000000000050214.

17. Wu M, Zhao P, Cui H. Content accuracy and reliability of myocarditis-related information on short-video platforms: A cross-sectional study of Bilibili and TikTok. Research Square. April 9, 2026. doi:10.21203/rs.3.rs-9257168/v1. (preprint)

18. Xu RH, Zhou L, Lu SY, et al,. Psychometric Validation and Cultural Adaptation of the Simplified Chinese eHealth Literacy Scale: Cross-Sectional Study. J Med Internet Res. 2020 Dec 7;22(12):e18613. doi: 10.2196/18613.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- Additionalfile2Supplementaryprotocol.docx