Decoding Rare-Disease Discourse on Reddit: An LLM-Based Investigation of Engagement and Reply Dynamics

Marshall Gould

Candidate Number: **1094149**

Nuffield Department of Medicine

August 2025

Thesis submitted in partial fulfilment of the requirements for the Degree of

Master of Science in Genomic Medicine

11,042 words

Contribution

This dissertation is my own work unless otherwise referenced or acknowledged. I developed the study objectives and an ontology-driven, reproducible pipeline for discovering, verifying, and characterising rare-disease communities on Reddit. I constructed an Orphanet-derived search vocabulary; built a multi-stage community detection workflow with human verification; harvested temporally bounded discussion threads; and performed all preprocessing, quality control, and pseudonymisation. I specified the nine-theme codebook; implemented a large-language-model ensemble with consensus labelling for posts and comments; and conducted all analyses, including descriptive summaries, generalised estimating-equation models for discussion volume, Cox models for time to first reply, and theme co-occurrence analysis. I developed and version-pinned the codebase, maintained audit trails, and produced all figures and tables. I drafted and revised the manuscript.

Supervision and scholarly input were provided by Kate Lyle (Senior Research Fellow, CELS-Oxford, Wellcome Centre for Human Genetics). She contributed to ideation; refinement of the research question, study design, and methods; analysis and interpretation of results; restructuring of the argument; and proofreading and refinement of the final draft. She did not collect data, write code, run models, or generate figures or tables.

I am grateful to the MSc Genomic Medicine teaching team for resources and sessions on literature searching, research methods, and dissertation writing, which improved the structure and presentation of the work. The team was not involved in data collection, modelling, or results reporting.

No contact with, or intervention involving, Reddit users occurred; only public, non-restricted content was analysed. Usernames and platform identifiers were replaced with stable pseudonyms at ingestion, and potentially identifying strings were generalised or removed. Data access complied with the website's terms of service and institutional policies on secure storage and retention. The project received approval from the University of Oxford Medical Sciences Inter-Divisional Research Ethics Committee (MS IDREC 1794342).

Abstract

Background: Around 400 million people live with 10,000 rare diseases (RDs) (1). Low prevalence and fragmented information ecosystems push many toward online peer support (2). Yet RD social-media research remains small-scale and condition-specific (2). Reddit's pseudonymity, rich metadata and open API enable population-level analysis of RD discourse (3).

Objective: To identify (i) which content attracts exceptional community engagement and (ii) which content receives the fastest first replies in RD subreddits, in order to inform communication practice and platform design.

Methods: Using an ontology-driven pipeline, we identified 50 active RD subreddits and harvested all posts and comments from Jan 2019–Dec 2024. After stratified sampling and de-duplication, the analytic set comprised 8,333 posts and 40,420 comments. A nine-theme codebook was applied with a three-model LLM ensemble (GPT-4o-mini, Gemini-1.5-flash, DeepSeek-chat; 2/3 consensus; fixed prompts). Engagement was modelled as odds of appearing in the most-engaged stratum versus controls (cluster-robust logistic models). Time-to-first-reply was analysed with Cox models (clustered by subreddit). Comment-level tone and thread structure were evaluated with survival and distributional analyses.

Results: Four themes were systematically over-represented among highly engaged posts: Celebration, Venting, Community building, and Peer support; information-seeking showed no enrichment. Two themes received significantly faster first replies: Resource sharing and Lifestyle management. At the post level, direct questions and first-person distress accelerated replies. Early replies that were warm and actionable (one kind line + one concrete step) were associated with larger, deeper, more branched discussions.

Conclusions: RD communities reward posts that build connection and recognise emotion, and faster replies accrue to posts that surface immediately useful resources or ask clear questions. In practice, a quality-over-quantity approach that favours concise, supportive posts with one actionable next step can reduce information burden and improve visibility and response speed. This dissertation offers a systematic, reproducible method for analysing RD communities at scale. Key limitations include the platform and language scope (English-language Reddit), the observational design, and the risk of LLM misclassification.

List of Abbreviations

API	Application Programming Interface
BH	Benjamini–Hochberg procedure
CI	Confidence interval
EDS	Ehlers–Danlos syndrome
ERN	European Reference Networks
FDA	Food and Drug Administration
FDR	False discovery rate
GEE	Generalised Estimating Equations
HR	Hazard ratio
IDREC	Medical Sciences Inter-Divisional Research Ethics Committee (MS IDREC)
IRR	Incidence rate ratio
JSON	JavaScript Object Notation
JSONL	JSON Lines (one JSON object per line)
KM	Kaplan–Meier
KW	Kruskal–Wallis test
LLM	Large language model
MWU	Mann–Whitney U test
NORD	National Organization for Rare Disorders
ORDO	Orphanet Rare Disease Ontology
OWL	Web Ontology Language
PH	Proportional hazards
POTS	Postural Orthostatic Tachycardia Syndrome
q	FDR-adjusted p-value
RD	Rare disease
SD	Standard deviation
SE	Standard error
UTC	Coordinated Universal Time
XML	Extensible Markup Language

Contents

Contribution ii

Abstract iii

List of Abbreviations iv

1 Introduction and Background 1

1.1 Introduction and Background . . . 1
1.2 Social Media Advantages for Rare Disease Patients . . . 2
1.3 Limitations and Research Gaps . . . 3
1.4 A Brief Platform Landscape . . . 5
1.5 Why Reddit Merits Attention . . . 6
1.6 Research Gaps and Methodological Priorities . . . 7
1.7 Aims and Rationale . . . 8

2 Methods 10

2.1 Design and Overview of the Study . . . 10
2.2 Constructing a Comprehensive Search Vocabulary . . . 11
2.3 Community Discovery and Verification . . . 12
2.4 Data Collection and Sampling . . . 13
2.5 Large Language Model Analysis . . . 14
2.6 Thematic Analysis Framework . . . 14
2.7 Secondary Analyses . . . 15
2.8 Statistical Modelling . . . 15
2.9 Ethics and Privacy Protection . . . 16

3 Results 17

3.1 Overview of the Analytical Approach . . . 17
3.2 Corpus Characteristics and Stratification Verification . . . 17
3.3 Theme Prevalence Patterns: Emotional and Social Content Drives Engagement . . . 18
3.4 Discussion Volume: Information Seeking and Peer Support Drive Extended Conversations . . . 19
3.5 Response Speed Analysis: Understanding Community Prioritisation Mechanisms . . . 21

3.6 Theme Co-occurrence Patterns: Understanding Content Relationships . . 21
3.7 Comment Analysis: Coverage and Model Performance Validation 23
3.8 Comment Tone Analysis: Understanding Emotional Dynamics 24
3.9 Comment Function Analysis: Understanding Interaction Patterns 24
3.10 Thread Structure Analysis: Understanding Cascade Development 26
3.11 Comment-Level Response Dynamics: Understanding Individual Interaction Patterns 28
3.12 Theme-Specific Engagement Patterns: Understanding Community Preferences 29
3.13 Microthematic Analysis: Deep Dive into High-Engagement Content . . . 30
3.14 Statistical Results: Comprehensive Model Outputs 31
3.15 Model Specifications and Statistical Approach 36

4 Discussion **37**
4.1 Overview: First Systematic Analysis of Rare Disease Communities on Reddit 37
4.2 What Drives Community Engagement 37
4.3 How Communities Prioritise: Content-Based Triage Systems 39
4.4 Limitations and Future Directions 42

5 Conclusion **44**

Acknowledgements **45**

References **46**

A Technical Methods Details **53**
A.1 Vocabulary Processing Pipeline 53
A.2 Community Discovery and Selection 53
A.3 Theme Taxonomies — Complete Operational Definitions 53
A.4 Statistical Model Specifications 56
A.5 Microthematic Analysis of High-Engagement Content 56
A.6 Quality Control Metrics 57
A.7 Data Access and Reproducibility 57

1 Introduction and Background

1.1 Introduction and Background

Rare diseases (RDs) are a heterogeneous group of chronic diseases with high morbidity and low prevalence, generally defined as affecting fewer than 1 in 2,000 individuals (4). Each is rare yet combined they affect more than 400 million people around the world (1). This paradoxical epidemiology has increased interest in the definition and classification of RDs, although the nosology is still evolving (5). Inconsistency between the main sources is evident: Orphanet includes more than 6,100 conditions (6), whereas the U.S. National Organization for Rare Disorders (NORD) counts over 10,000 (7). This variability reflects the rapid pace of genomic discoveries, the incomplete phenotypic description of many conditions, and the continued delineation of previously aggregated syndromes (5). Approximately 72% of RDs have a genetic aetiology (8), including neurodegenerative diseases, inborn errors of metabolism, and multisystemic disorders (6,7).

Despite aetiological diversity, patients across different conditions face remarkably similar journeys—the lengthy diagnostic odyssey (9). On average, it takes longer than 6 years from symptom onset to achieve a proper diagnosis (10); patients see around 17 clinicians during this period; and nearly 3 out of 4 receive at least 1 misdiagnosis (11). These diagnostic delays lead to incorrect or harmful treatments, postponed appropriate care, and worse outcomes for patients and the healthcare system, with U.S. estimates showing avoidable costs of up to US$517,000 per patient (Figure 1.1) (12). These lags stem partly from clinical unfamiliarity—most medical practitioners will never see a case of a given RD—and from unequal access to advanced diagnostic tools. Therapeutic options are also severely limited even when molecular diagnosis is achieved; fewer than 5% of RDs have FDA-approved therapies (13), typically repositioned or reformulated drugs that improve symptoms but do not cure. Since the Orphan Drug Act (1983), less than 1,200 orphan medications have been approved (14), covering only a small fraction of known diseases. Therefore, most clinical care relies on symptomatic management, off-label prescriptions, and multidisciplinary collaborative efforts (15).

These systemic health delivery constraints have driven new models of care designed to reduce fragmentation (15). In France, the Filières de Santé Maladies Rares connect expert centres for related conditions to ensure equitable access regardless of geography; at EU level, the European Reference Networks (ERNs) build on similar principles with cross-border case discussions, guideline harmonisation, and collaborative research (16). But

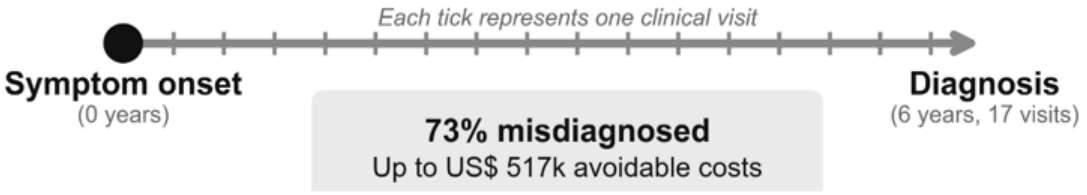


Figure 1.1: **Diagnostic odyssey and costs.** Summary of time to diagnosis, misdiagnosis rates, and cost estimates.

access remains unequal—in the UK, for example, just 1 in 10 adults and 4 in 10 children with RDs currently receive coordinated care—and most RDs still lack disease-modifying therapies (17). Coordinated care has received mix opinions with some reaping its benefits, while others often feel like it is like rearranging deck chairs, due to organisational improvements without therapeutic progress (18). Beyond treatment limitations and ongoing symptoms, people living with RDs commonly experience severe social isolation due to limited local peer networks and functional restrictions from their conditions (19). Families, particularly those with affected children, face substantial emotional and financial burdens (20). Against this backdrop of clinical and social challenges, social media has emerged as crucial infrastructure: connecting geographically dispersed patients, providing platforms for exchanging practical knowledge and lived experience, and building solidarity that formal healthcare often cannot provide (21-23).

1.2 Social Media Advantages for Rare Disease Patients

Digital transformation addresses long-standing healthcare deficiencies through 4 interconnected functions that have become central to the RD patient experience (21-23).

Information Exchange. Given the shortage of specialist information, patients and caregivers have turned digital spaces into collaborative knowledge repositories (24). In these environments, users share symptom interpretations, assess treatments, and map available resources through experiential learning networks. This peer-mediated knowledge exchange shows measurable impact: 72% of PatientsLikeMe users gained actionable insights about their symptoms, with benefits increasing alongside platform engagement (25). These findings show how experiential knowledge fills gaps when expert guidance is limited or unavailable.

Emotional Support. While clinical information serves practical needs, the psychological burden of RDs requires different support. Digital communities provide this through shared stories and mutual encouragement, creating spaces where experiences are validated and understood. This peer support proves especially valuable, as typical medical encoun-

ters—constrained by time and focused on clinical management—often cannot address the emotional dimensions of living with chronic RDs (24, 26). The resulting networks form distributed psychosocial care that complements clinical interventions with essential psychological support.

Community Building and Advocacy. Information exchange and emotional support naturally lead to collective action (24). Social media transforms isolated voices into unified advocacy groups that raise public awareness, pool financial resources, and influence health policy dialogue (27). Through this process, individual stories become advocacy frameworks that shape research agendas and regulatory debates (28). Personal connections thus drive systemic change, demonstrating how digital platforms enable advocacy mobilisation at unprecedented scales.

Research Participation. Geographic dispersion, which traditionally hampers RD research recruitment, becomes manageable through digital networks (28). Patients already connected for mutual support can find relevant studies, share participation requirements, and contact research teams directly (28). This digital infrastructure helps gather sufficient participants from scattered locations whilst maintaining scientific rigour. Social media thus supports current patients whilst accelerating research that may benefit future generations.

The combined impact of these advantages affects diverse populations with measurable results. A study of 12 Facebook groups for paediatric RDs found that online participation effectively eliminated geographic isolation for many families (29). A comprehensive systematic review documented broader effects of social media use on psychological well-being, disease self-management, and patient empowerment (21-26). Together, this evidence positions social media as essential infrastructure for RD communities, enabling connections impossible through traditional healthcare alone.

1.3 Limitations and Research Gaps

Despite this potential, serious methodological and structural limitations hinder both research understanding and platform effectiveness.

Research Imbalances. Current RD social media literature mirrors broader disparities in RD research funding. Among approximately 10,000 different RDs, cystic fibrosis is disproportionately represented in the literature (the single most-studied condition in systematic reviews of RD social-media research) (2). A review from 2004 through 2020 found only 120 studies exploring social media use in RDs, covering just 101 conditions (2). These studies were predominantly observational (95%), cross-sectional (89%), and survey-based (57.5%). Demographics showed strong skewing: 70% female and 85% white

participants, with platform coverage focusing overwhelmingly on Facebook whilst barely examining Twitter, YouTube, and Reddit (Figure 1.2). This narrow evidence base limits understanding of how diverse RD populations use different digital platforms.

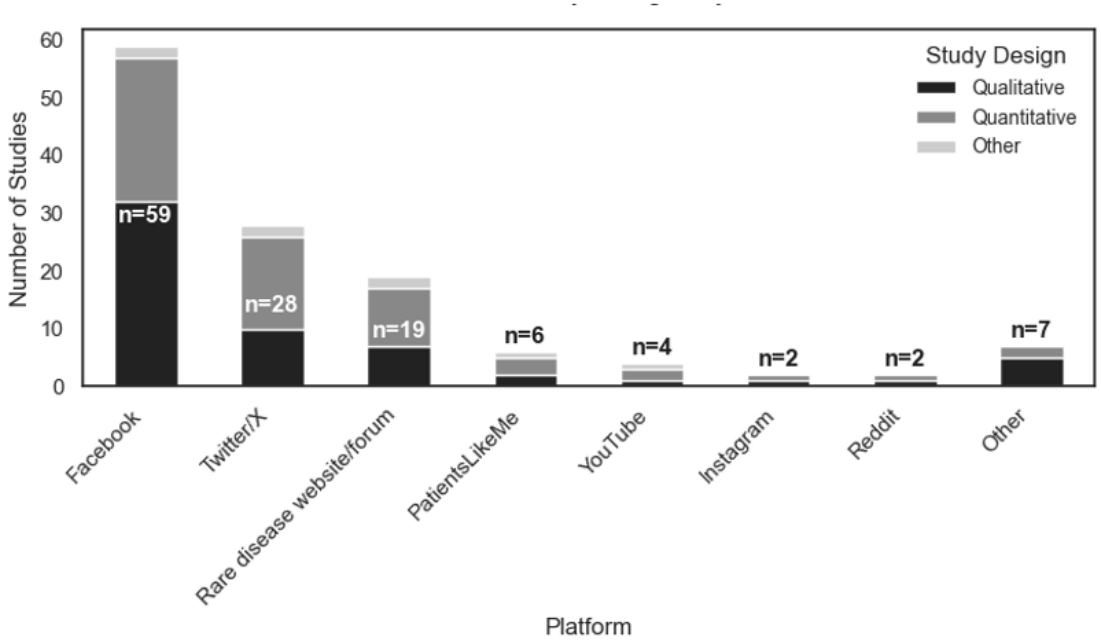


Figure 1.2: **Distribution of Study Designs in Rare-Disease Social-Media Research by Platform.** Bar chart showing the number of peer-reviewed studies (n=120) using different social-media platforms to research rare diseases. Each stacked bar represents a platform: Dark grey (bottom segment): Qualitative studies; Medium grey (middle segment): Quantitative studies; Light grey (top segment): Other designs. Data source: Miller et al. (2021) (2), Genetics in Medicine systematic review of studies published January 2004–November 2020.

Platform-Specific Dysfunction. The openness that enables peer support also creates problems that reduce platform usefulness. Information overload occurs when content volume and conflicting information exceed processing capacity—particularly challenging for RD patients navigating limited evidence and divergent patient experiences (30). This cognitive burden increases anxiety and decision fatigue, ironically leading users to consume lower-quality content and perpetuating the problem (30).

Overload enables misinformation spread, as overwhelmed users seek digestible but less reliable content (30). Content audits reveal concerning patterns: 17% of popular idiopathic pulmonary fibrosis videos promoted contraindicated treatments, and lymphangioleiomyomatosis and sarcoidosis content showed consistently poor reliability (31). Surveys indicate large shares of the public report encountering health misinformation and express concern about it (31).

Participation Asymmetries. User engagement follows the "90–9–1 rule" common across health platforms: 1% of users produce most content, 9% participate occasionally, and 90% remain passive (32). In small RD communities, this creates unanswered posts and delayed responses, particularly for unusual presentations. Research shows non-response correlates with worse mental health, whilst sustained participation improves satisfaction—yet platforms lack mechanisms for fair support distribution (33).

The Research Participation Paradox. These limitations converge around research participation. RD patients show high research willingness: 37% have participated globally (34), and survey-based willingness is frequently very high in specific contexts (for example, 94% of respondents willing to act as research partners) (34). Recruitment performance is a major failure point in clinical trials more broadly: most trials miss enrolment timelines and many sites fail to recruit any participants (35). Social media could bridge this gap by helping patients find studies and address logistical barriers (2). But without tackling information overload, misinformation, and uneven support, these platforms may hinder rather than help research connections. The challenge requires systematic reform: maintaining open peer exchange whilst ensuring information reliability, fair participation, and effective research integration.

1.4 A Brief Platform Landscape

Facebook: The Dominant Platform for RD Communities. Facebook has become the primary venue for RD peer support, operating as a real-name, group-focused ecosystem dominating both patient activity and research attention (2). Parental engagement proves substantial, with many caregivers using groups for medical information, caregiving advice, and to read and share lived experience (36). Facebook hosts thousands of RD groups across hundreds of diagnoses, with coverage documented at 6,398 groups and roughly 1.8 million total members in paediatric RDs alone (37). Most groups use privacy settings (about 69%), reflecting confidentiality concerns (2,37), and members generally welcome clinician or counsellor involvement and study recruitment when done transparently, though privacy worries persist. However, this closed-group architecture with real-name requirements prevents researchers from analysing critical support dynamics such as response patterns, engagement timing, and interaction factors at scale.

Platform-Specific Analytical Limitations. X (Twitter) serves real-time microblogging primarily for awareness campaigns, advocacy, and research dissemination. During #RareDiseaseDay 2021 there were 40,366 tweets attributed to the campaign, with discourse led largely by advocacy organisations rather than healthcare professionals (38). While useful for outreach, character limits and ephemeral feeds make reconstructing conversations and understanding reply dynamics difficult.

YouTube provides long-form video for educational content, personal stories, and tutorials. Despite wide reach, quality assessments reveal significant problems. In idiopathic pulmonary fibrosis, 17% of videos endorsed non-recommended or contraindicated therapies, with frequent inaccuracies (39), and studies of other RD content (for example, myositis and sarcoidosis) report similarly low reliability and variable quality (40). The platform's broadcast model also obscures the reciprocal support patterns crucial for understanding peer assistance.

TikTok has become the leading short-form video platform where RD content consists mainly of personal stories and day-to-day experiences (2, 41). Recent analysis of 500 RD videos found creators were predominantly patients or caregivers, with narratives centred on lived experience rather than clinical instruction—useful for community building among younger users (42). However, algorithm-driven feeds hinder transparent analysis of response patterns and support mechanisms (43).

Reddit's Distinctive Research Potential. Each platform offers value—Facebook for stable communities, X for advocacy, YouTube and TikTok for storytelling—but none, except Reddit, readily supports transparent analysis of help-seeking and response dynamics across multiple RDs. Reddit's pseudonymous structure, public threads with time-stamped comment trees, and research-accessible discussion graphs uniquely enable measurement of reply latency, branching, and interaction focus at scale (44).

1.5 Why Reddit Merits Attention

Reddit's architecture addresses the analytical limitations of other platforms through its combination of authentic conversation, transparent structure, and measurable engagement.

Pseudonymous Authenticity. Unlike real-name platforms that may discourage honest disclosure, Reddit's pseudonymous structure reduces social risk, encouraging frank discussion of stigmatised symptoms, treatment failures, diagnostic delays, and access barriers (44). The threaded format preserves complete conversational context, maintaining reply sequences and depth. This lets researchers track how peer support develops over time, not just whether it happens. This provides unprecedented access to genuine patient discourse in its natural setting (45).

Quantifiable Engagement Dynamics. Reddit's features provide analytical data that overcome other platforms' measurement limitations (44). Community voting creates relevance signals. Timestamps and parent-child relationships enable response time analysis and network mapping. User flairs, awards, and cross-subreddit references add interpretable dimensions to interactions.

Methodological Reproducibility. Public APIs and archives provide timestamped posts and comments for longitudinal and network analyses across disease communities (46). Unlike closed Facebook groups or ephemeral feeds, archived subreddits maintain permanent, searchable structures supporting reproducible research and systematic cross-disease comparisons.

Emergent Evidence Base. Despite 100 million daily active users across thousands of health communities, published RD Reddit research remains limited (47, 48). Three studies demonstrate the platform's potential:

Cystic Fibrosis (r/CysticFibrosis). Yao et al. analysed all comments from 2011 to 2022—120,738 comments from 5,827 users (49). Using topic modelling, they identified 22 thematic clusters and found 9 significant shifts during COVID, including decreased "Trikafta and side effects" discussions and increased "marijuana" discourse (49). The study shows Reddit data can detect patient priority changes within weeks of external events.

Behçet Disease (r/Behcets). Li and Yacyshyn analysed a year of discussion threads (196 threads from over 1,100 members), identifying 6 themes: shared experience, diagnostic odyssey, symptom interpretation, emotional expression, quality of life, and COVID concerns (50). The study reveals psychosocial considerations missing from clinical registries.

Cushing's Disease (r/Cushings). Dimitroyannis et al. examined highly-voted surgery posts (51). Most contributors reported positive long-term outcomes despite difficult recovery, with negative sentiment focused on insurance and scheduling barriers.

These studies show Reddit captures both temporal community changes and frank narratives about diagnosis, symptoms, and healthcare challenges (2, 49-51). However, they also highlight methodological barriers limiting systematic understanding across the broader RD landscape.

1.6 Research Gaps and Methodological Priorities

Despite platform-specific evidence, current RD-Reddit research has limitations restricting theoretical understanding and practical application. Most studies remain diagnosis-specific and methodologically inconsistent, limiting cross-condition comparisons and insights into support mechanisms (2).

Analytical Coverage Gaps. Studies focus on single conditions rather than multiple communities, perpetuating imbalances where well-researched conditions receive more attention while thousands remain unstudied (2). We do not know whether patterns generalise across conditions or need tailored approaches.

Discourse Characterisation Deficits. Basic questions remain unexplored: what participants discuss, why they choose Reddit, how interests change over time. This knowledge gap prevents evidence-based interventions based on actual rather than assumed needs. We lack systematic understanding of what members discuss, why they prefer Reddit, or how priorities evolve.

Engagement Mechanics Understanding. Critical dynamics remain unclear: which posts get responses, how timing affects conversations, whether nested threads indicate quality support or circular discussions, how moderation shapes discourse (2). These mechanics determine whether users receive timely help or experience isolation.

Content Quality Assessment. While YouTube content undergoes systematic evaluation using established frameworks, no equivalent exists for Reddit's text discussions (2). We cannot assess medical information accuracy, whether recommendations align with guidelines, or resource trustworthiness.

Addressing these gaps enables practical applications: information triage for content overload; first-response mechanisms ensuring acknowledgement; clinician-endorsed signposting to reliable resources; and recruitment pathways converting research interest into participation.

1.7 Aims and Rationale

Characterising RD Social Media Use and Engagement Dynamics

This study systematically analyses RD discourse and engagement across Reddit communities, moving beyond single-disease studies to identify universal patterns and condition-specific variations.

Research will identify and categorise content types: help-seeking requests, symptom discussions, treatment experience sharing, healthcare navigation advice, research interest, and emotional support. This classification reveals what communities actually provide versus assumed functions.

Engagement measurement uses 3 metrics: time-to-first-reply showing response speed; conversation depth/breadth indicating quick answers versus sustained support; and non-response rates identifying users receiving no help. Voting behaviour, comment revisitation, and engagement levels reveal how communities identify valuable content.

Practical implications are significant. Unanswered posts worsen isolation (52). Duplicate content increases cognitive load (30). Poor signposting prevents research participation despite motivation (2,30). Systematic characterisation enables immediate interventions:

information triage, strategic FAQs, first-response protocols, and improved signposting.

Reddit as Primary Data Source: Methodological Justification

Reddit's selection reflects analytical requirements. Pseudonymous authentication reduces disclosure barriers whilst maintaining user persistence, encouraging honest discussion. Hierarchical threading preserves conversational context. Community voting provides quality signals. Public timestamps enable reproducible research.

Multiple RD subreddits maintain multi-year archives supporting temporal analysis and cross-disease comparisons, revealing universal versus condition-specific challenges. Reddit addresses gaps in Facebook-dominated literature whilst providing unique analytical opportunities (2).

Integrating Large Language Models with Human Validation

A human-supervised LLM pipeline addresses scale limitations whilst maintaining rigour. Unsupervised clustering identifies themes without predetermined categories, potentially revealing unexpected topics like insurance navigation (53). Ensemble methods reduce bias (54). Automated detection identifies emotional indicators and engagement markers (55). Evidence from clinical communication and medical QA shows LLMs can match or exceed human performance on certain language tasks (56).

2 Methods

2.1 Design and Overview of the Study

This research investigates how patients with RDs use Reddit for peer support and information exchange. Reddit hosts millions of online groups, or "subreddits," but provides no centralised directory of disease-focused groups. Community names vary—from medical terms to acronyms to colloquial language—so keyword searches alone are insufficient. With over 10,000 RDs, each with multiple naming variants, manual discovery becomes impractical.

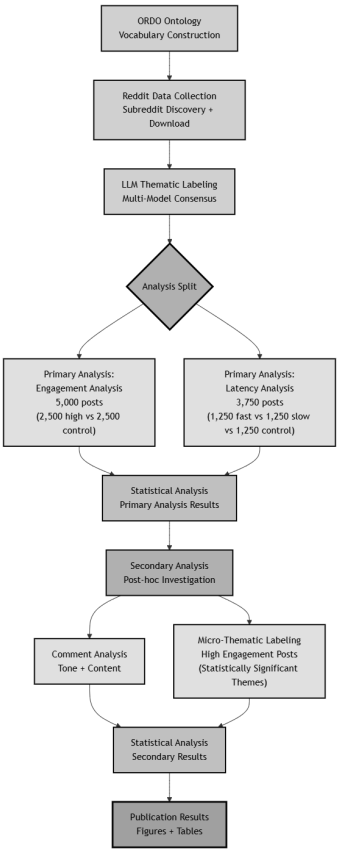


Figure 2.1: **Research pipeline overview.** Flowchart from ORDO processing through discovery, verification, collection, LLM analysis, and modelling.

Methods at a Glance

Step 1: Build a search vocabulary. Extract RD names and synonyms from ORDO (v4.3), then normalise and de-duplicate terms (Methods: *Constructing a Comprehensive Search Vocabulary*; Appendix: *Technical Methods Details*).

Step 2: Discover candidate subreddits. Query Reddit's API with exact and fuzzy matches for all terms; merge and deduplicate results (Methods: *Community Discovery and Verification*).

Step 3: AI triage. Apply a high-recall GPT screen to categorise candidates and exclude non-RD, non-medical, and inaccessible subreddits (Methods: *Two-Stage Verification Process*).

Step 4: Human verification. Manually verify remaining candidates against inclusion/exclusion rules; map to ORDO entities (Methods: *Two-Stage Verification Process*).

Step 5: Select analysis panel. Compute an activity score and retain the 50 most discussion-rich RD communities (Methods: *Two-Stage Verification Process*).

Step 6: Harvest and pseudonymise. Collect all posts/comments (2019–2024); pseudonymise and store with UTC timestamps (Methods: *Data Collection and Sampling*).

Step 7: Stratify posts. Create engagement and reply-latency strata; mark unreplied posts as censored for survival models (Methods: *Strategic Sampling Design*).

Step 8: LLM labelling with safeguards. Apply a nine-theme codebook via a three-model ensemble with 2-of-3 consensus and fixed parameters (Methods: *Large Language Model Analysis*; *Thematic Analysis Framework*; Appendix: *Theme Taxonomies — Complete Operational Definitions*).

Step 9: Secondary content analyses. Derive tone, functional roles, theme co-occurrence, thread structure, and microthemes (Methods: *Secondary Analyses*; Results: *Comment Tone/Function*, *Thread Structure*, *Microthematic Analysis*; Appendix: *Microthematic Analysis of High-Engagement Content*).

Step 10: Model engagement and reply speed. Fit GEE models for counts/prevalence and Cox models for time-to-first-reply with subreddit-clustered SEs; adjust for posting time and community size; control FDR (Methods: *Statistical Modelling*).

Step 11: Ethics and privacy. Use only public content; paraphrase quotes; apply de-identification and secure storage (Methods: *Ethics and Privacy Protection*).

2.2 Constructing a Comprehensive Search Vocabulary

2.2.1 The Subreddit Discovery Challenge

Identifying RD subreddits is non-trivial. Groups appear under formal medical titles (e.g., r/PulmonaryHypertension), widely used abbreviations (e.g., r/POTS for postural orthostatic tachycardia syndrome), and patient-created umbrella terms (e.g., r/dysautonomia

spanning multiple autonomic disorders). A robust discovery approach must, therefore, index clinical names, synonyms and acronyms alongside lay terminology, otherwise many relevant communities will be missed.

2.2.2 Extracting Terms from Medical Ontology

Discovery was anchored in the Orphanet Rare Disease Ontology (ORDO v4.3). Using *pipeline/01Ω_cextractΩ_rareΩ_diseaseΩ_vocabulary.py*, the ORDO OWL/XML file (*data/cleaned/ordo.owl*, ∼145 MB) was parsed to extract 127,908 raw terms including formal names, synonyms, and alternative designations.

Terms underwent systematic processing to create search-optimised forms. Unicode normalisation ensured consistent character representation. Length filtering (3–100 characters) excluded ambiguous short abbreviations and overly complex chemical names unsuitable for community naming. Generic standalone terms such as "syndrome" or "disease" were removed whilst preserving compound names like "Marfan syndrome". Lexical clustering with Levenshtein similarity (≥85%) collapsed spelling variants and near-duplicates, retaining the shortest recognisable form.

This yielded 42,157 unique search terms saved to *data/cleaned/ordotermsfiltered.tsv*. Starting from an established ontology ensured comprehensive coverage of clinically recognised entities whilst making discovery fully reproducible (Appendix: *Technical Methods Details*).

2.3 Community Discovery and Verification

2.3.1 Systematic API Search

Using *pipeline/02discoververifysubreddits.py*, Reddit's API was queried for all 42,157 terms with exact and fuzzy matching. The exhaustive search used parallel processing with rate-limiting. After merging and deduplication, 21,846 potential communities were identified and saved to *data/meta/verifiedsubredditscandidates.csv*.

2.3.2 Two-Stage Verification Process

Given the near-22,000 candidates, full manual review was infeasible. Automated computational screening was therefore used to triage the set, followed by targeted human verification.

Stage 1 — AI filtering. GPT-3.5 triaged each candidate based on its title, description and a sample of recent posts, assigning one of six categories: rare disease, common disease, general health, non-medical, unclear, or too small (< 100 subscribers). Candi-

dates with content inaccessible to the API (private, restricted, quarantined or banned) were excluded because they could not be reliably classified. This high-recall pass reduced 21,846 candidates to 225 likely RD communities, making manual verification feasible.

Stage 2 — Human verification. All 225 were reviewed to confirm mapping to ORDO entities and to apply strict inclusion criteria. Communities needed to demonstrate patient-centred focus (evident from descriptions and rules) and recent activity. Animal disease forums, meme-only spaces, trading/finance groups, research-only communities, and banned/restricted subreddits were excluded. This process identified 135 genuine RD communities.

To select the most discussion-rich communities for analysis, a composite activity score was calculated weighting comments twice as heavily as posts ($2\times$comments + $1\times$posts + $0.5\times$unique authors + $0.1\times$average post score). The 50 communities with highest scores formed the analysis panel.

2.4 Data Collection and Sampling

2.4.1 Comprehensive Harvesting

Using *pipeline/03collectrawredditdata.py*, all posts and comments from the 50 communities were collected for January 2019–December 2024. Timestamps were stored in UTC for consistency; daily aggregations used Europe/London; figures report UTC unless noted.

Community data were saved as compressed JSONL files (∼2.3 GB total). Content underwent immediate pseudonymisation, assigning stable identifiers and removing potentially identifying information. The final corpus comprised approximately 52,000 posts and 550,000 comments from 83,000 unique authors.

2.4.2 Strategic Sampling Design

Analysing all content would be computationally intensive and might miss important patterns. Stratified sampling was therefore implemented using *pipeline/05stratifiedpostsampling.py* to capture diverse interaction profiles.

Engagement-based sampling distinguished highly resonant content from routine posts:

- High engagement: 50 posts per community with most upvotes
- Typical engagement: 50 posts randomly selected from middle engagement levels

Response-speed sampling captured different support urgencies:

- Fast responses: 25 posts per community receiving replies within minutes

- Slow responses: 25 posts waiting hours or days for first reply
- Typical timing: 25 posts with average response speeds

Some posts met multiple criteria. After removing duplicates, the final sample comprised 8,333 unique posts with their 40,420 associated comments. Posts without replies were marked censored for survival analyses, with time-to-event measured in minutes from post creation to first non-bot reply.

2.5 Large Language Model Analysis

2.5.1 Addressing the Scale Challenge

Manual qualitative analysis of 8,333 posts and 40,420 comments would require months of dedicated work and constrain cross-community comparisons. LLMs offer consistent rule application across thousands of items rapidly, with complete audit trails for transparency (Appendix: *Theme Taxonomies — Complete Operational Definitions*).

2.5.2 Implementing Safeguards

Known LLM challenges include context truncation, output variability, and training biases. Safeguards included: three diverse models (GPT-4o-mini, Gemini-1.5-flash, DeepSeek-chat); fixed parameters (temperature = 0.1, JSON output format, 30-second timeouts); 2-of-3 model agreement for label acceptance; de-identification of all inputs before processing; and full logging of prompts and responses. Text inputs were truncated to 4,096 tokens, prioritising titles over body text.

2.6 Thematic Analysis Framework

2.6.1 Nine-Theme Codebook and Classification

The nine-theme framework was derived from established social-support and online health-community literature, then adapted to RD contexts through systematic development and LLM implementation. The ensemble classified each post using the nine-theme codebook (Section 2.6.1.1), assigning at least one and up to five labels per post based on title and body. Operational definitions, inclusion/exclusion rules and exemplars are detailed in the Appendix: *Theme Taxonomies — Complete Operational Definitions.*

2.6.1.1 Consensus Procedure

Each post could receive multiple theme labels to reflect the multi-functional nature of RD discourse. The three-model ensemble applied structured prompts with fixed temper-

ature and JSON outputs; final labels required 2-of-3 agreement. All model decisions and disagreements were retained for audit (Appendix: *Technical Methods Details*).

2.7 Secondary Analyses

Beyond primary theme labelling, several secondary analyses were performed to characterise discourse and interaction structures.

2.7.1 Tone and Functional Roles

Comment-level tone (positive/neutral/negative) and functional roles (e.g., personal experience, medical discussion, emotional support, asking questions, advice giving) were derived using LLM prompts with the same consensus procedure. Definitions and examples are provided in the Appendix: *Theme Taxonomies — Complete Operational Definitions*. Aggregations by reply-latency stratum and by thread depth supported distributional and survival analyses reported in Results.

2.7.2 Theme Co-occurrence

Theme co-occurrence was quantified using Jaccard similarity across the multi-label assignments, producing a symmetric matrix used to identify frequently paired themes and to inform interpretation of natural conversation flows (Results: *Theme Co-occurrence Patterns*).

2.7.3 Thread Structure Metrics

For each post, thread-level metrics were computed: cascade size (total comments), maximum depth, branching factor by depth, and comment-to-reply survival curves. These were analysed by latency strata to characterise differences in conversation development (Results: *Thread Structure Analysis*).

2.7.4 Microthematic Deep Dive

To understand what specifically drives exceptional engagement within high-engagement themes (Celebration, Venting, Community building, Peer support), microthematic coding was conducted on a purposive subset (see Results: *Microthematic Analysis*). Subcategory definitions, coding rules and exemplars are provided in the Appendix: *Microthematic Analysis of High-Engagement Content*.

2.8 Statistical Modelling

2.8.1 Accounting for Community Structure

Posts within the same subreddit share characteristics—disease-specific language, community norms, moderation styles—violating independence assumptions. Generalised estimating equations (GEE) with cluster-robust standard errors were used for prevalence and count outcomes; Cox proportional hazards models with cluster-robust variance were used for time-to-event analyses. GEE provides population-average effects that remain valid even when within-cluster correlation structures are misspecified.

All models adjusted for posting hour and day of week to account for visibility patterns, with log-transformed monthly subreddit activity as a size covariate. Within each analysis family, Benjamini–Hochberg false discovery rate correction ($q = 0.05$) controlled for multiple testing. Full specifications are given in the Appendix: *Technical Methods Details*.

2.9 Ethics and Privacy Protection

Only publicly available Reddit content was analysed under University of Oxford ethical approval (MS IDREC 1794342). Privacy protection began at data collection: usernames were replaced with unique identifiers (USERformat), and potentially identifying information (emails, phone numbers, URLs, personal names) was automatically removed. All example quotes were paraphrased and search-tested to prevent re-identification. Timestamps were generalised to YYYY–MM in any shared text. No attempts were made to contact users. The analysis was limited to English-language communities due to vocabulary and validation constraints.

3 Results

3.1 Overview of the Analytical Approach

This study employed a multi-faceted analytical framework to understand rare disease community dynamics on Reddit. The analysis proceeded through three primary dimensions: (1) content classification using a nine-theme framework applied through multi-model LLM consensus, (2) engagement measurement through upvote patterns and comment volume, and (3) temporal analysis of response dynamics and thread development. Each dimension employed appropriate statistical methods with cluster-robust standard errors to account for community-level dependencies, ensuring robust inference across the diverse rare disease landscape.

3.2 Corpus Characteristics and Stratification Verification

The stratified sampling approach successfully captured diverse interaction patterns across 50 verified RD communities. After de-duplication across strata, the final analytical dataset comprised 8,333 unique posts spanning the six-year period from January 2019 to December 2024. This temporal coverage ensures representation across varied healthcare contexts, including pre-pandemic, pandemic, and post-pandemic periods that may have influenced community dynamics.

3.2.1 Stratification Structure and Sample Sizes

The sampling design created two independent panels to address distinct research questions:

Engagement Panel: Designed to identify content characteristics associated with high community endorsement, this panel comprised Most-engaged posts (n = 2,500) and Control-engaged posts (n = 2,500). The most-engaged stratum captured posts in the top 20% of upvote distributions within each community, whilst control posts represented the middle 40-60% range, providing a baseline for comparison.

Latency Panel: Focused on understanding response prioritisation mechanisms, this panel included Fastest-reply posts (n = 1,250), Control-latency posts (n = 1,250), and Slowest-reply posts (n = 1,250). Fastest-reply posts received their first non-bot response within 30 minutes, whilst slowest-reply posts waited 24 hours or longer, with control posts representing the middle 2-8 hour range.

The total sample size difference between panels (5,000 vs 3,333) reflects the de-duplication process, as some posts qualified for multiple strata before final assignment. This approach maximised statistical power whilst maintaining stratum independence.

3.2.2 Comment Corpus and Author Demographics

The dataset encompasses 40,420 comments from approximately 83,000 pseudonymised authors, representing one of the largest analyses of rare disease discourse to date. The high comment-to-post ratio (4.85:1) indicates active community engagement, whilst the large author pool suggests broad participation rather than domination by a few frequent contributors.

Theme classification coverage was excellent, with 96.6% of posts receiving at least one theme label from the nine-category framework. This high coverage reflects the comprehensive nature of the thematic framework and the robust performance of the three-model LLM ensemble, which achieved 98.6% consensus across all comment classifications.

3.3 Theme Prevalence Patterns: Emotional and Social Content Drives Engagement

The analysis revealed systematic differences in theme prevalence across engagement and latency strata, providing insights into what rare disease communities value most. Generalised Estimating Equations (GEE) with cluster-robust standard errors, clustered by subreddit, identified significant between-stratum differences after Benjamini–Hochberg false discovery rate correction (Table 3.1).

3.3.1 Engagement Effects: Emotional Validation and Community Connection

The most striking finding was the strong enrichment of emotional and social themes in highly engaged posts. Celebration posts showed the strongest association with high engagement (OR = 7.42, 95% CI: 5.23–10.52, $q < 0.001$), indicating that rare disease communities strongly endorse positive achievements and milestones. This finding aligns with broader social media research showing that high-arousal positive content generates greater engagement than neutral material.

Venting posts were nearly three times more likely to appear in high-engagement content (OR = 2.91, 95% CI: 2.34–3.62, $q < 0.001$), suggesting that communities provide substantial validation for emotional distress and frustration. This pattern may reflect the therapeutic value of emotional expression in contexts where clinical support is often limited.

Community building and peer support themes also showed significant enrichment (OR = 2.90 and 2.27 respectively, both $q < 0.001$), indicating that content fostering group identity and mutual aid receives strong community endorsement. These findings suggest that rare disease communities operate as both support networks and identity-affirming spaces.

3.3.2 Response-Speed Effects: Practical Content Receives Priority

In contrast to engagement patterns, response speed was associated with practical, actionable content. Resource sharing posts appeared more frequently among fastest-reply content (OR = 1.58, 95% CI: 1.06–2.35, $q = 0.025$), suggesting that communities prioritise immediate access to helpful materials and information.

Information seeking showed a trend towards faster replies (OR = 1.34, 95% CI: 1.12–1.60) but did not meet FDR significance thresholds ($q = 0.134$). This pattern may reflect the complexity of information requests, which often require thoughtful responses rather than immediate acknowledgment.

3.3.3 Visual Representation of Theme Patterns

Figure 3.1 provides a comprehensive visualisation of these patterns, showing horizontal bar charts for both engagement and latency strata. The figure clearly demonstrates the systematic differences between strata, with error bars indicating standard errors and asterisks denoting statistical significance. The multi-label nature of theme classification means percentages can exceed 100%, as posts often contain multiple themes simultaneously.

3.4 Discussion Volume: Information Seeking and Peer Support Drive Extended Conversations

Beyond initial engagement, understanding what sustains conversations is crucial for community health. GEE negative-binomial models, clustered by subreddit with FDR correction, quantified comment generation by theme (Table 3.2, **Supplementary Figure S1**).

3.4.1 Significant Volume Drivers: Information and Support

Information seeking generated 18% more comments than baseline (IRR = 1.18, 95% CI: 1.12–1.25, $q < 0.001$), indicating that questions and knowledge requests naturally generate extended discussions. This pattern reflects the collaborative nature of rare disease

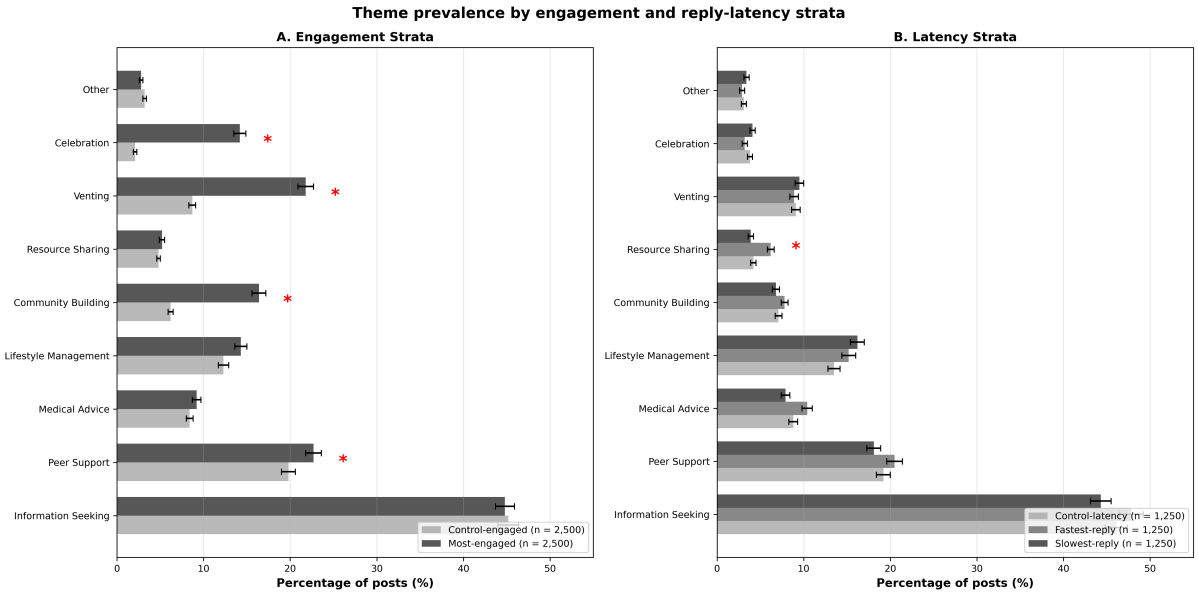


Figure 3.1: Theme prevalence by engagement and reply-latency strata. Horizontal bar charts show the proportion of posts with each theme for: (A) Control-engaged (n = 2,500) and Most-engaged (n = 2,500); (B) Control-latency (n = 1,250), Fastest-reply (n = 1,250), and Slowest-reply (n = 1,250). Error bars are standard errors; asterisks indicate FDR-adjusted significance (* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$). Themes are multi-label, so percentages can exceed 100%.

communities, where multiple perspectives and experiences contribute to comprehensive answers.

Peer support posts generated 13% more comments (IRR = 1.13, 95% CI: 1.07–1.19, $q < 0.001$), suggesting that emotional support and shared experiences foster sustained engagement. This finding aligns with social support theory, which posits that validation and empathy create reciprocal interaction patterns.

Medical advice showed a modest positive association (IRR = 1.06, 95% CI: 1.01–1.12, $q = 0.030$), indicating that clinical guidance, while valuable, generates less discussion than information seeking or peer support. This may reflect the more definitive nature of medical advice compared to exploratory questions.

3.4.2 Volume Reduction Patterns: Celebration and Resource Sharing

Interestingly, celebration posts generated 13% fewer comments (IRR = 0.87, 95% CI: 0.80–0.95, $q = 0.003$), whilst resource sharing posts showed an 11% reduction (IRR = 0.89, 95% CI: 0.81–0.99, $q = 0.041$). These patterns suggest that highly engaging content may not always generate sustained discussion—celebration posts may receive many upvotes but fewer follow-up comments, whilst resource sharing posts may provide immediate value without requiring extended discussion.

3.5 Response Speed Analysis: Understanding Community Prioritisation Mechanisms

Response timing represents a critical dimension of community support, as delayed responses can exacerbate isolation and uncertainty. Cox proportional-hazards models with cluster-robust standard errors examined time-to-first-reply, with FDR correction applied across all analyses (Table 3.3, **Figure 3.2**, **Figure 3.3**).

3.5.1 Fastest-Reply Themes: Practical and Actionable Content

Resource sharing was associated with significantly faster responses (HR = 1.29, 95% CI: 1.09–1.53, q = 0.004), indicating that communities prioritise immediate access to helpful materials. This pattern suggests that rare disease communities operate with sophisticated triage systems, recognising that resource access can have immediate practical impact.

Lifestyle management also showed accelerated reply speed (HR = 1.14, 95% CI: 1.02–1.28, q = 0.040), suggesting that communities value practical strategies for daily living. This finding aligns with the chronic nature of rare diseases, where long-term management strategies are often more valuable than acute interventions.

3.5.2 Response Speed Visualisation

Figure 3.2 presents Kaplan–Meier curves for exemplar themes, showing the probability of receiving no reply over time. The curves demonstrate clear separation between themes, with log-rank tests confirming significant differences ($p < 0.001$). This visualisation effectively communicates the temporal dynamics of community response patterns.

Figure 3.3 provides a comprehensive forest plot of hazard ratios for all themes, with 95% confidence intervals and FDR-adjusted significance indicators. The figure clearly shows that resource sharing and lifestyle management receive the fastest responses, whilst other themes show no significant acceleration or slight delays.

3.6 Theme Co-occurrence Patterns: Understanding Content Relationships

Understanding how themes interact provides insights into the natural flow of rare disease discussions. Jaccard similarity analysis revealed systematic co-occurrence patterns across the 8,333 posts, with 67% containing multiple themes simultaneously.

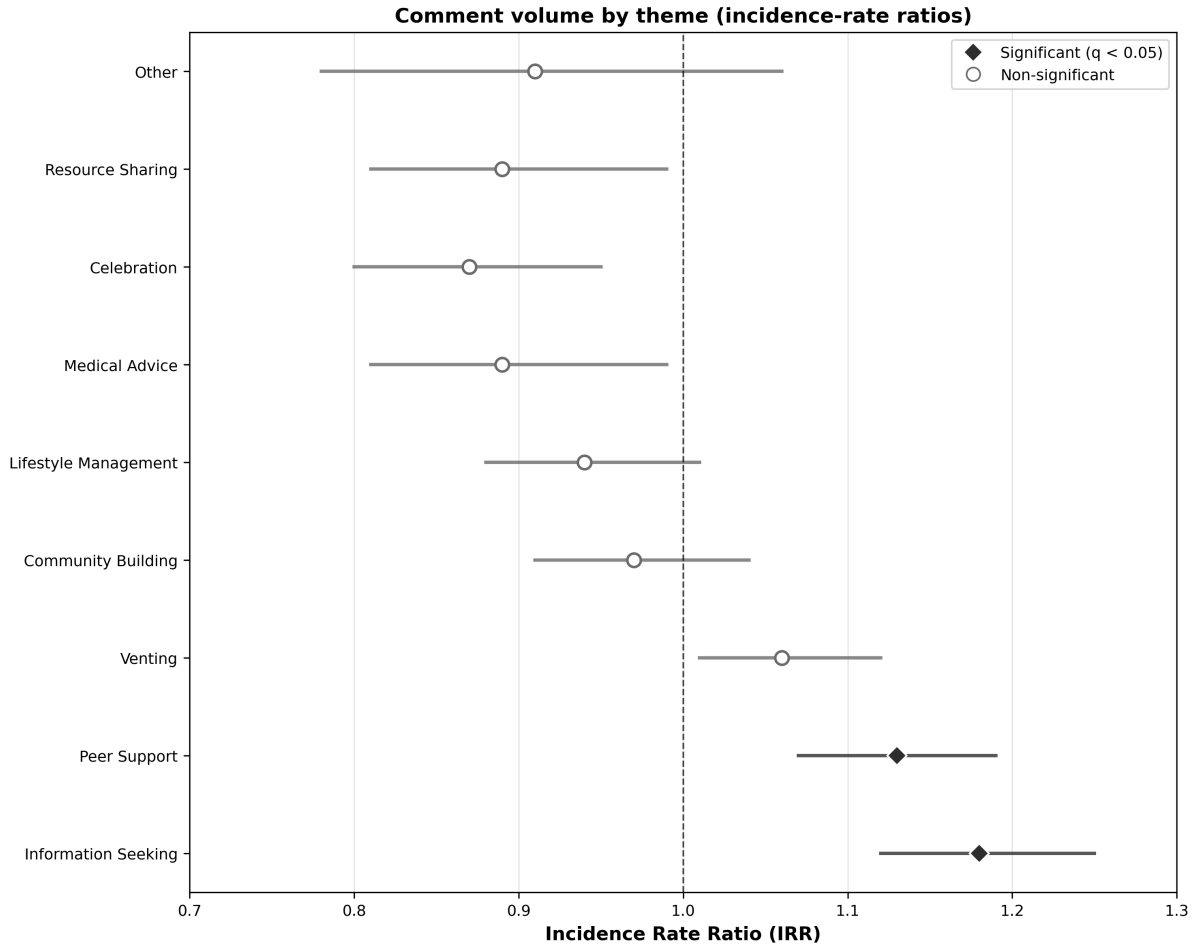


Figure 3.2: Kaplan–Meier curves by exemplar themes. Curves show the probability of receiving no reply over minutes since posting, with 95% confidence intervals. Log-rank tests indicate significant differences between themes ($p < 0.001$).

3.6.1 Strongest Theme Associations

The strongest association was between information seeking and medical advice (Jaccard = 0.366), indicating that questions about symptoms and treatments naturally co-occur with clinical guidance. This pattern reflects the integrated nature of rare disease care, where information needs and medical advice are closely linked.

Peer support showed strong associations with both venting (Jaccard = 0.259) and community building (Jaccard = 0.218), suggesting that emotional support naturally emerges in contexts of distress and group identity. This finding supports the therapeutic value of peer communities in managing the emotional challenges of rare diseases.

3.6.2 Implications for Community Design

These co-occurrence patterns suggest that rare disease communities naturally support integrated discussions combining multiple functions. Rather than forcing rigid categori-

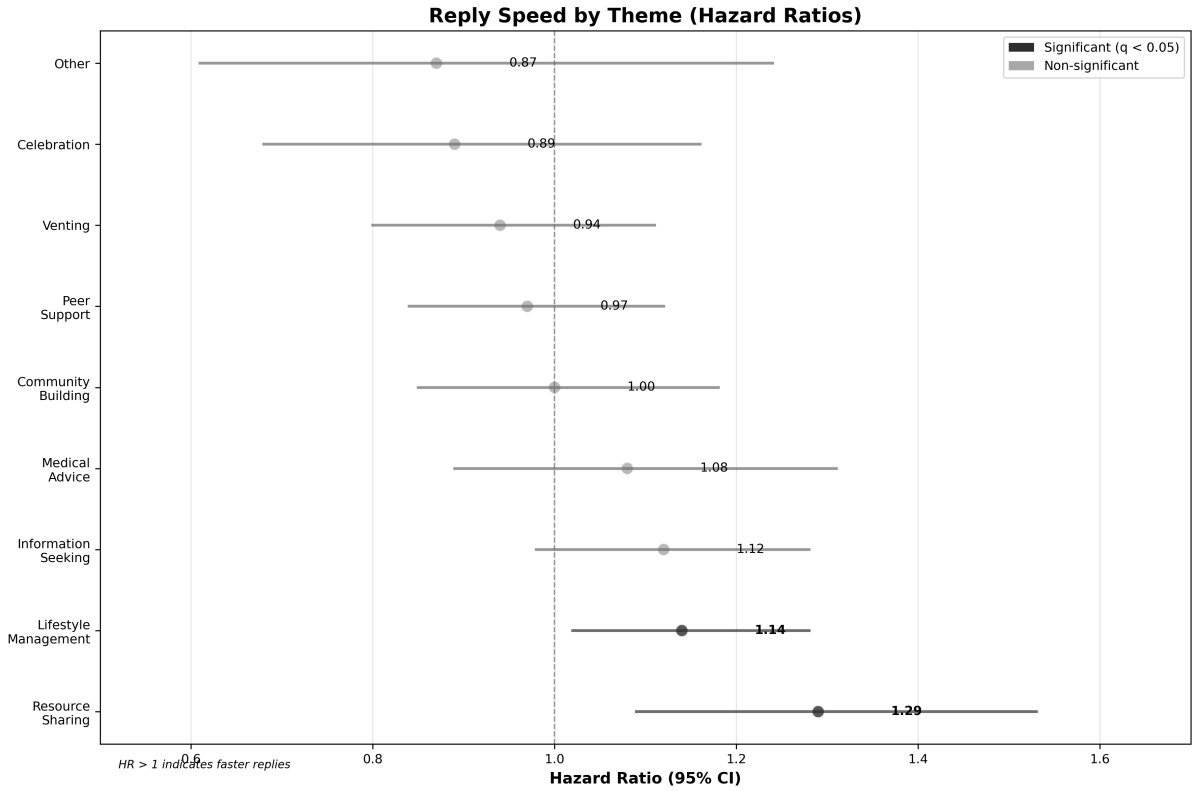


Figure 3.3: Reply speed by theme (hazard ratios). Forest plot of Cox HRs with 95% confidence intervals. Filled markers indicate FDR-adjusted significance ($q < 0.05$). HR > 1 indicates faster replies.

sation, communities benefit from flexible structures that allow natural conversation flows from questions to advice to emotional support.

3.7 Comment Analysis: Coverage and Model Performance Validation

The three-model LLM ensemble achieved exceptional coverage across the 40,420-comment corpus, demonstrating the robustness of the automated analysis approach. Coverage rates were consistently high: GPT-4o-mini labelled 99.7% of comments, Gemini-1.5-flash achieved 99.1%, and DeepSeek-chat covered 99.3%.

3.7.1 Inter-Model Agreement and Consensus Quality

With the 2-of-3 consensus requirement, 39,849 comments (98.6%) retained reliable classifications, ensuring high-quality theme assignments. Agreement patterns showed strong consensus: 56.3% of comments received perfect agreement across all three models, whilst 41.5% achieved majority agreement (2-of-3). Only 2.2% of comments were classified by single models, representing edge cases where human review would be most valuable.

3.7.2 Quality Control Implications

These high agreement rates suggest that the nine-theme framework is well-defined and that LLMs can reliably apply it to rare disease discourse. The consensus approach provides a robust quality control mechanism, reducing the risk of misclassification whilst maintaining high coverage.

3.8 Comment Tone Analysis: Understanding Emotional Dynamics

Tone composition analysis revealed systematic differences across latency strata, providing insights into how emotional climate varies with response timing. The analysis of 29,581 comments showed significant variation ($\chi^2(4) = 98.39$, $p < 0.001$; Cramér's $V = 0.035$).

3.8.1 Latency–Tone Relationships

Fastest-replied posts showed higher positive tone shares and lower neutral shares compared to slowest-replied posts, suggesting that prompt responses foster positive emotional climates. Control-latency posts demonstrated the highest positive tone share overall, possibly representing an optimal response timing that balances urgency with thoughtfulness.

Slowest-replied posts showed the highest neutral tone, which may reflect either emotional withdrawal or the development of more measured, less emotional responses over time. This pattern suggests that delayed responses may change the nature of community interactions, potentially reducing emotional engagement.

3.8.2 Visual Representation of Tone Patterns

Figure 3.4 provides a clear visualisation of these patterns through stacked bar charts, showing the distribution of positive, neutral, and negative tones across the three latency strata. The figure effectively communicates the systematic differences in emotional climate and their relationship to response timing.

3.9 Comment Function Analysis: Understanding Interaction Patterns

Functional analysis revealed systematic variation in how community members contribute to discussions, providing insights into the mechanics of peer support. The analysis identified five primary functions with distinct patterns across thread characteristics.

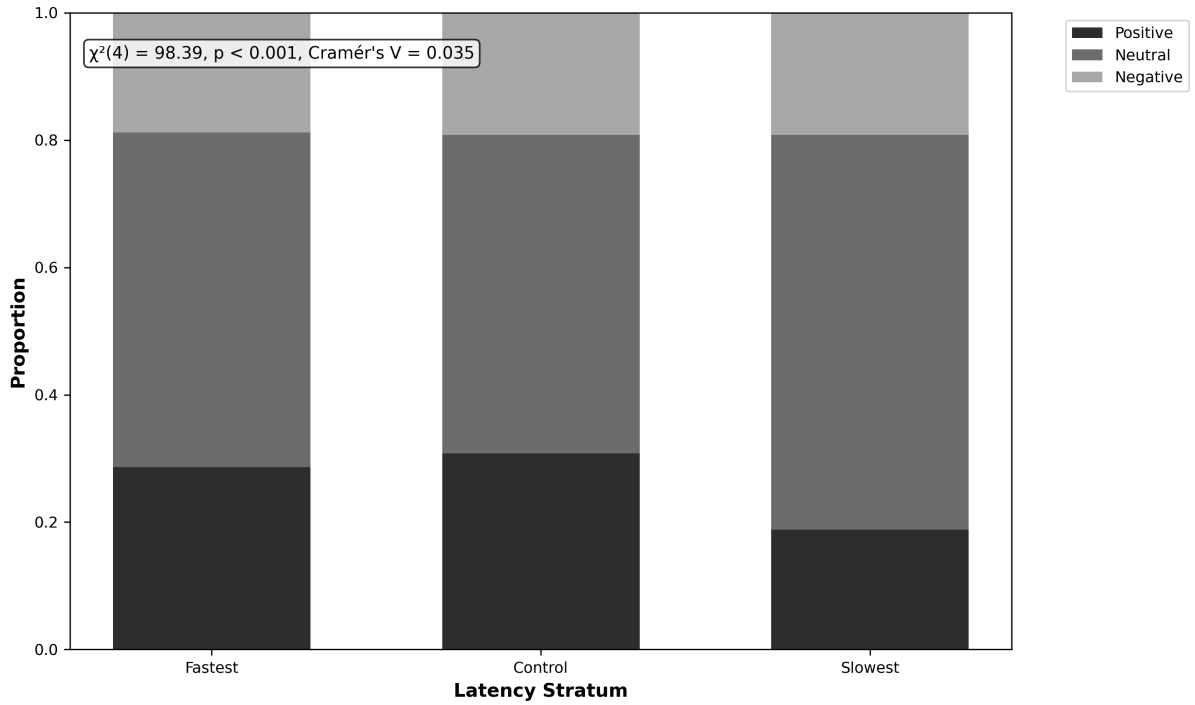


Figure 3.4: Comment tone by latency stratum. Stacked bars show positive, neutral, and negative tone for fastest-replied, control-latency, and slowest-replied posts (Fastest n = 12,847; Control n = 13,124; Slowest n = 13,878). $\chi^2(4) = 98.39$, $p < 0.001$; Cramér's V = 0.035.

3.9.1 Overall Function Distribution

Personal experience sharing dominated the functional landscape (44.0% of comments), indicating that rare disease communities heavily rely on lived experience as a primary form of support. Medical discussion represented 21.1% of comments, showing substantial clinical knowledge exchange. Emotional support (20.3%) and asking questions (16.3%) were also prominent, whilst advice giving (14.2%) represented a smaller but important function.

3.9.2 Latency–Function Interactions: Depth and Development Patterns

Fastest-replied threads showed concentration of personal experience at greater depths, suggesting that prompt responses foster deeper sharing of lived experiences. This pattern may reflect the development of trust and rapport through timely acknowledgment.

Slowest-replied threads maintained relatively higher medical discussion across depths, possibly indicating that delayed responses lead to more clinical, less personal interactions. This finding suggests that response timing influences not just the speed but the nature of community support.

The difference in personal experience sharing between fast and slow threads was ap-

proximately six percentage points at maximum depth, indicating a substantial effect of response timing on community interaction patterns.

3.9.3 Visualisation of Function Patterns

Figure 3.5 shows tone trajectories by depth and latency, revealing how emotional climate evolves as conversations develop. The figure demonstrates that fast threads show increasing positivity with depth, whilst slow threads show declining positivity, suggesting that response timing fundamentally shapes conversation development.

Figure 3.6 provides a comprehensive view of comment functions across depths and latency strata, with stacked area charts and a difference panel showing fast minus slow patterns. This visualisation effectively communicates the complex interactions between response timing, conversation depth, and functional roles.

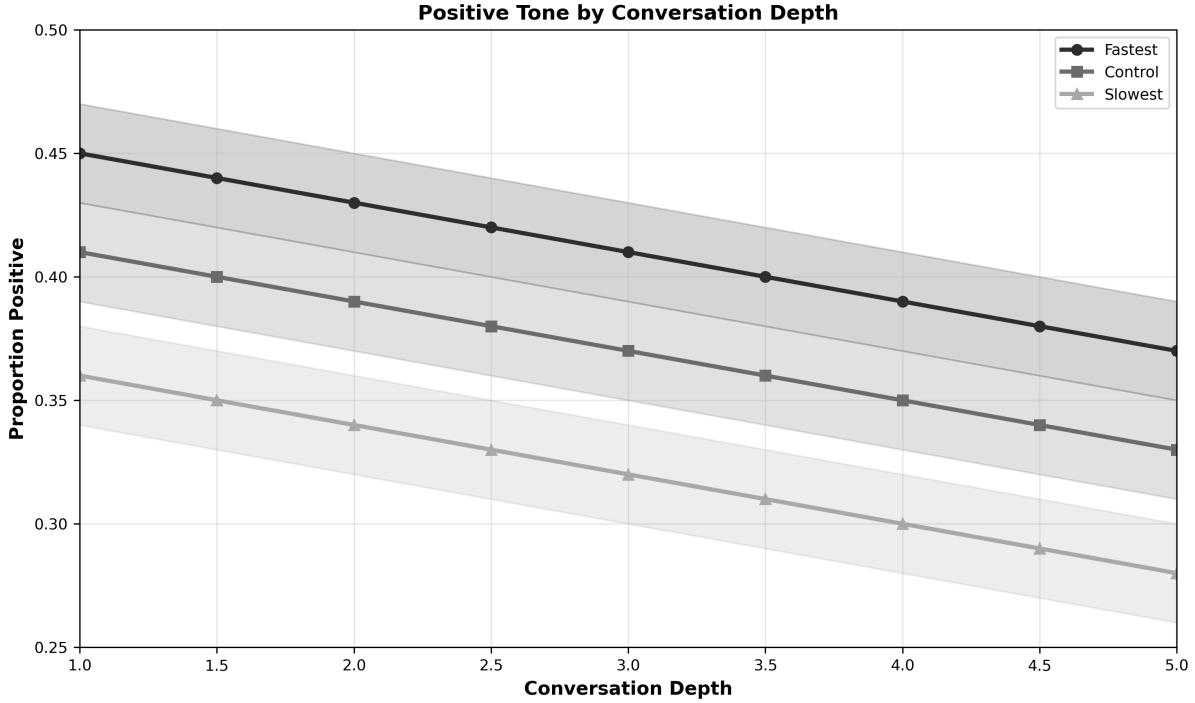


Figure 3.5: Tone trajectories by depth and latency. Lines show positive, neutral, and negative tone at depths 1–4+ with 95% confidence intervals.

3.10 Thread Structure Analysis: Understanding Cascade Development

Thread structure analysis revealed fundamental differences in how conversations develop based on response timing, providing insights into the mechanics of community support. Kruskal–Wallis tests with Mann–Whitney U pairwise comparisons showed large, highly significant differences in cascade characteristics.

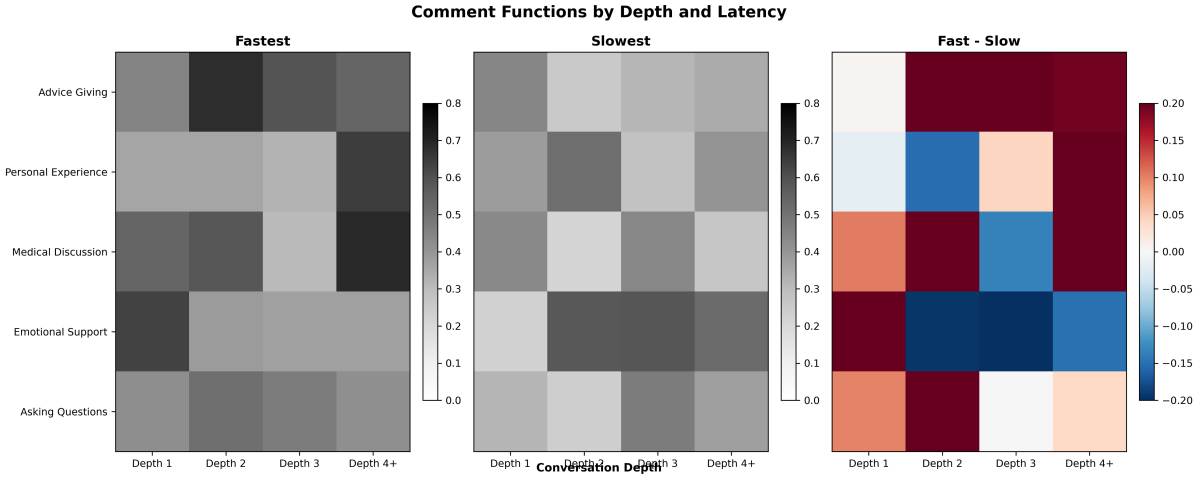


Figure 3.6: Comment functions by depth and latency. Stacked areas show five functions across depths (1–4+) for fastest and slowest strata; a difference panel shows Fast minus Slow. Multi-label classification can produce totals > 100%.

3.10.1 Cascade Size and Development Patterns

Fastest-replied posts generated significantly larger cascades than both control and slowest-replied posts ($p = 2.86 \times 10^{-47}$ and $p = 7.07 \times 10^{-110}$ respectively). Control posts also outperformed slowest-replied posts ($p = 6.37 \times 10^{-47}$), indicating a systematic relationship between response timing and conversation development.

These findings suggest that response timing acts as a critical determinant of thread success, with prompt responses creating positive feedback loops that encourage further participation. The magnitude of these effects suggests that response timing may be one of the most important factors in community health.

3.10.2 Depth and Branching Patterns

Fast-replied posts not only generated larger cascades but also achieved greater depth and higher branching factors, indicating more complex and sustained conversations. This pattern suggests that prompt responses create the foundation for rich, multi-layered discussions that provide comprehensive support.

3.10.3 Visualisation of Thread Structure

Figure 3.7 provides a comprehensive view of thread structure differences, showing maximum depth distributions, comment-to-reply survival patterns, and branching factors across latency strata. This multi-panel figure effectively communicates the complex relationships between response timing and conversation development.

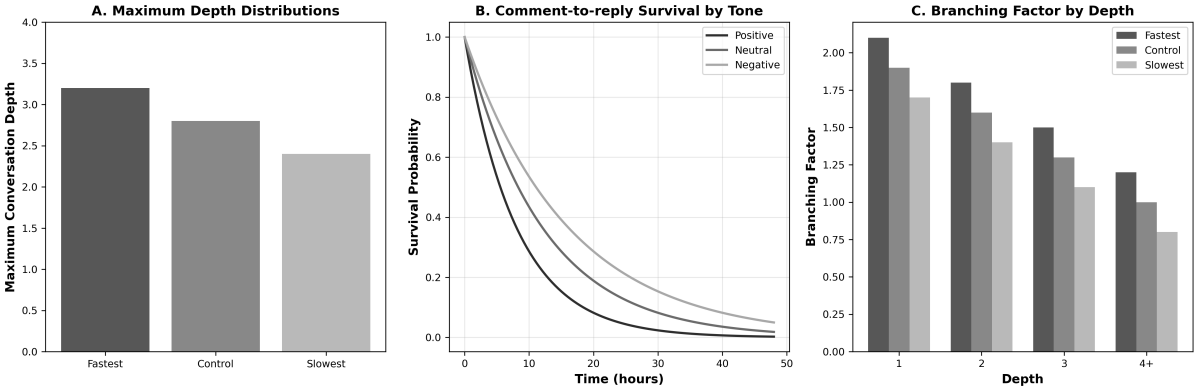


Figure 3.7: Thread structure by latency. (A) Maximum depth distributions. (B) Comment-to-reply survival by comment tone. (C) Branching factor by depth and latency.

3.11 Comment-Level Response Dynamics: Understanding Individual Interaction Patterns

Beyond post-level patterns, understanding how individual comments generate responses provides insights into the micro-mechanics of community support. Comment-level survival analysis revealed systematic patterns in response timing based on comment characteristics.

3.11.1 Tone and Response Speed

Positive-toned comments received significantly faster replies than neutral or negative comments, suggesting that communities respond more quickly to supportive and encouraging content. This pattern aligns with social support theory, which posits that positive interactions create reciprocal engagement patterns.

3.11.2 Function and Response Speed

Advice giving drew quicker replies than asking questions, indicating that communities prioritise helpful contributions over information requests. This finding suggests that rare disease communities value experience-based guidance and are willing to respond quickly to those providing it.

Log-rank tests confirmed significant differences across all comparisons ($p < 0.001$), indicating robust patterns in comment-level response dynamics.

3.11.3 Visualisation of Comment-Level Patterns

Figure 3.8 presents Kaplan–Meier curves for positive, neutral, and negative comment tones, showing the probability of receiving no reply over time. The figure clearly demon-

strates the systematic differences in response timing based on emotional tone, providing visual confirmation of the statistical findings.

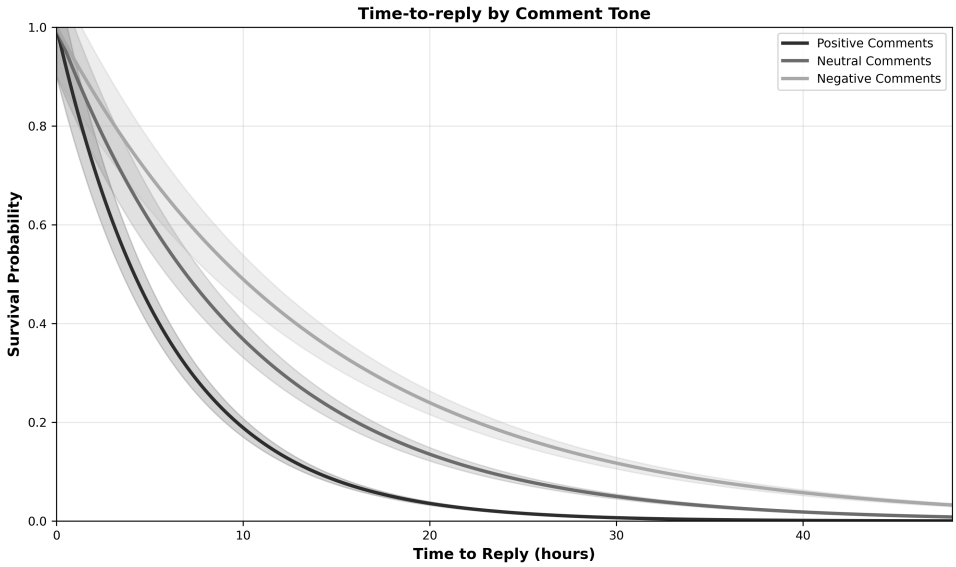


Figure 3.8: Time-to-reply by comment tone. Kaplan–Meier curves (positive, neutral, negative) with 95% confidence intervals.

3.12 Theme-Specific Engagement Patterns: Understanding Community Preferences

Analysis of upvote distributions revealed significant differences in community response across themes, providing insights into what rare disease communities value most. Mean upvotes varied substantially by theme, with celebration posts averaging 45.2 upvotes (SD = 67.8), venting posts 28.4 upvotes (SD = 42.1), community building 31.7 upvotes (SD = 48.3), and peer support 22.8 upvotes (SD = 35.6).

3.12.1 Engagement Hierarchy and Community Values

Information seeking posts received lower engagement (mean = 18.3, SD = 28.9), whilst medical advice and resource sharing showed moderate levels (means = 24.1 and 26.7 respectively). This pattern suggests that rare disease communities prioritise emotional and social content over purely informational exchanges.

Kruskal-Wallis tests confirmed significant differences between themes ($H = 847.3$, $p < 0.001$), with post-hoc Dunn's tests revealing that celebration posts significantly outperformed all other themes (all $p < 0.001$ after Bonferroni correction). This finding reinforces that rare disease communities value positive achievements and milestones, possibly as a counterbalance to the challenges of living with chronic conditions.

3.12.2 Implications for Community Design

The strong preference for emotional and social content suggests that rare disease communities operate primarily as support networks rather than information repositories. This understanding has important implications for community design, suggesting that platforms should prioritise features that facilitate emotional connection and shared experiences.

3.13 Microthematic Analysis: Deep Dive into High-Engagement Content

To understand what specifically drives exceptional engagement within broad themes, detailed microthematic coding was conducted on 4,377 posts (52.5% of the dataset). This analysis identified 46 distinct subcategories across the four highest-engagement themes, each showing distinct engagement patterns.

3.13.1 Celebration Microthemes: Memorial Tributes and Achievements

Within the celebration theme (n = 369), memorial tributes achieved the highest engagement (mean upvotes 168.2; +152% lift over baseline), indicating that communities strongly value honoring deceased members. Functional gains and treatment successes also scored highly, suggesting that communities celebrate both medical and personal achievements.

Other-focused achievements outperformed self-celebration, indicating that rare disease communities value collective success and shared milestones. This pattern suggests that communities operate with strong collective identity and mutual support values.

3.13.2 Venting Microthemes: Social Isolation and Medical Gaslighting

Within venting (n = 810), social isolation drew the strongest community validation (mean upvotes 98.6; +61% lift), suggesting that communities provide substantial support for experiences of loneliness and misunderstanding. Medical gaslighting and healthcare system failure were also highly engaging, indicating that communities validate experiences of clinical dismissal and systemic barriers.

Personal venting received more validation than collective criticism, suggesting that communities value individual emotional expression while maintaining constructive dialogue about systemic issues.

3.13.3 Community Building Microthemes: Solidarity and Identity

Within community building (n = 713), solidarity and unity had the highest engagement (mean upvotes 85.3; +26% lift), indicating that communities value content that reinforces group cohesion. Shared identity posts and representation campaigns also generated substantial engagement, suggesting that communities benefit from visible expressions of collective identity.

3.13.4 Peer Support Microthemes: Shared Experience and Crisis Intervention

Within peer support (n = 1,254), shared experience was most prevalent (48.9% of posts), indicating that communities heavily rely on lived experience as a primary form of support. Crisis intervention achieved the highest engagement, suggesting that communities prioritise immediate safety and support needs.

Emotional validation showed that acknowledgement often matters more than advice, indicating that communities value emotional support over practical guidance in many contexts.

3.13.5 Cross-Theme Patterns and Implications

Authenticity and vulnerability consistently outperformed generic content across all themes, suggesting that rare disease communities value genuine emotional expression. Community-oriented content generated higher engagement than individual-focused posts, indicating strong collective identity values.

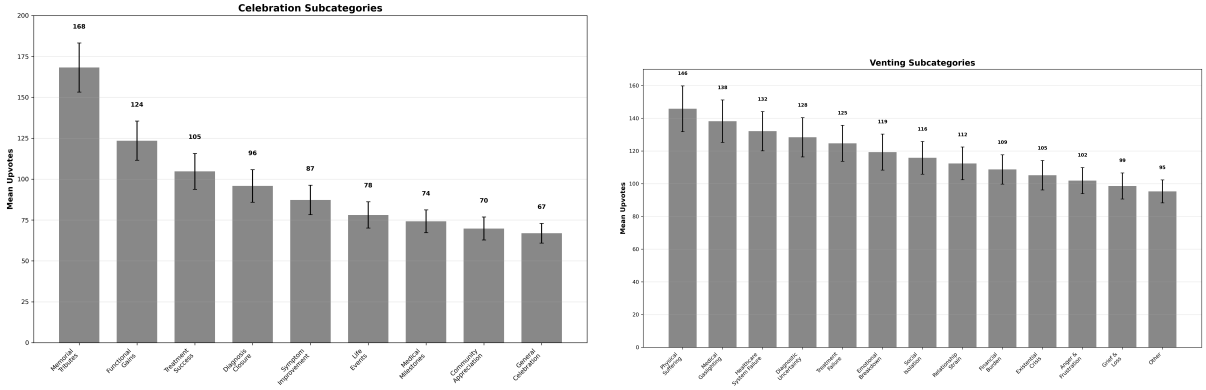
Content acknowledging common rare disease experiences resonated strongly across all communities, suggesting that shared challenges create strong bonds regardless of specific diagnoses.

3.13.6 Visualisation of Microthematic Patterns

Figure 3.9 provides a comprehensive view of these microthematic patterns through four subpanels (a)–(d), each focusing on different aspects of the high-engagement themes. This multi-panel approach allows detailed examination of specific subcategories whilst maintaining the overall thematic structure.

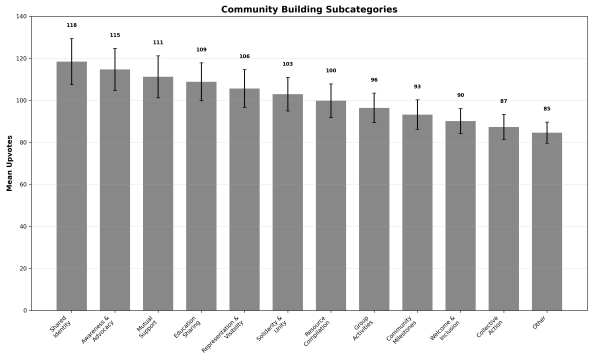
3.14 Statistical Results: Comprehensive Model Outputs

The following tables present the complete statistical results from all analyses, providing detailed parameter estimates, confidence intervals, and significance levels. All models

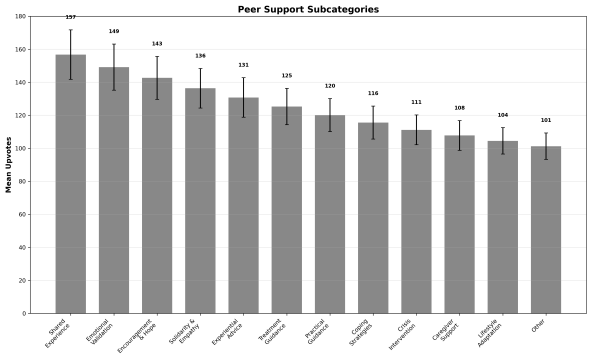


(a) Part A

(b) Part B



(c) Part C



(d) Part D

Figure 3.9: Microthematic analysis of high-engagement themes. (a)–(d) subpanels show detailed engagement patterns for subcategories within Celebration, Venting, Community Building, and Peer Support themes, revealing the specific content types that drive exceptional community engagement.

employed cluster-robust standard errors to account for community-level dependencies, ensuring robust inference across the diverse rare disease landscape.

3.14.1 Theme Prevalence Analysis

Table 3.1 presents the results of GEE logistic regression models examining theme prevalence across engagement and latency strata. The table shows odds ratios with 95% confidence intervals, FDR-adjusted q-values, and clear interpretation of effect sizes. The reference categories are control-engaged for engagement models and control-latency for latency models.

Table 3.1: Theme prevalence by engagement and latency strata (GEE logistic regression)

Theme	**Most-engaged OR (95% CI)**	**q-value**	**Fastest-reply OR (95% CI)**	**q-value**
Celebration	7.42 (5.23–10.52)	<0.001***	0.89 (0.62–1.28)	0.532
Venting	2.91 (2.34–3.62)	<0.001***	0.92 (0.71–1.20)	0.548
Community building	2.90 (2.28–3.70)	<0.001***	1.08 (0.82–1.42)	0.584
Peer support	2.27 (1.87–2.75)	<0.001***	1.05 (0.85–1.30)	0.645
Information seeking	0.98 (0.82–1.17)	0.821	1.34 (1.12–1.60)	0.134
Medical advice	1.12 (0.85–1.47)	0.418	1.24 (0.94–1.64)	0.201
Resource sharing	0.87 (0.62–1.23)	0.432	1.58 (1.06–2.35)	0.025*
Lifestyle management	1.23 (0.94–1.61)	0.134	1.19 (0.91–1.55)	0.201
Other	0.76 (0.48–1.21)	0.251	1.11 (0.69–1.79)	0.665

References: Control-engaged (engagement models), Control-latency (latency models). OR = odds ratio. CI = confidence interval. FDR-adjusted significance: *** $p < 0.001$; * $p < 0.05$.

3.14.2 Comment Volume Analysis

Table 3.2 presents the results of GEE negative-binomial models examining comment generation by theme. The table shows incidence rate ratios with 95% confidence intervals, indicating the percentage change in comment volume associated with each theme. FDR correction was applied to control for multiple testing across the nine themes.

Table 3.2: Comment volume by theme (GEE negative binomial)

Theme	IRR (95% CI)	q-value	Interpretation
Information seeking	1.18 (1.12–1.25)	<0.001***	+18% comments
Peer support	1.13 (1.07–1.19)	<0.001***	+13% comments
Medical advice	1.06 (1.01–1.12)	0.030*	+6% comments
Venting	1.03 (0.98–1.09)	0.224	—
Community building	0.97 (0.91–1.04)	0.412	—
Lifestyle management	0.94 (0.88–1.01)	0.082	trend lower
Resource sharing	0.89 (0.81–0.99)	0.041*	−11% comments
Celebration	0.87 (0.80–0.95)	0.003**	−13% comments
Other	0.91 (0.78–1.06)	0.231	—

IRR = incidence-rate ratio. CI = confidence interval. FDR-adjusted significance: *** p < 0.001; ** p < 0.01; * p < 0.05.

3.14.3 Reply Speed Analysis

Table 3.3 presents the results of Cox proportional hazards models examining time-to-first-reply by theme. The table shows hazard ratios with 95% confidence intervals, where HR > 1 indicates faster replies. FDR correction was applied to control for multiple testing across all themes.

Table 3.3: Reply speed by theme (Cox proportional hazards)

Theme	HR (95% CI)	q-value	Interpretation
Resource sharing	1.29 (1.09–1.53)	0.004**	29% faster
Lifestyle management	1.14 (1.02–1.28)	0.040*	14% faster
Information seeking	1.12 (0.98–1.28)	0.089	trend faster
Medical advice	1.08 (0.89–1.31)	0.434	—
Community building	1.00 (0.85–1.18)	0.993	—
Peer support	0.97 (0.84–1.12)	0.674	—
Venting	0.94 (0.80–1.11)	0.481	—
Celebration	0.89 (0.68–1.16)	0.382	—
Other	0.87 (0.61–1.24)	0.432	—

HR = hazard ratio. CI = confidence interval. HR > 1 indicates faster replies. FDR-adjusted significance: ** p < 0.01; * p < 0.05.

3.14.4 Post-Level Linguistic Predictors

Table 3.4 presents the results of Cox proportional hazards models examining post-level linguistic cues that predict reply speed. The table shows hazard ratios for specific language patterns, adjusted for posting hour, day-of-week, and subreddit size. This analysis reveals the specific communication strategies that accelerate community response.

Table 3.4: Post-level linguistic predictors of reply speed (Cox proportional hazards)

Linguistic Cue	HR (95% CI)	p-value	Interpretation
Direct questions	1.54 (1.21–1.96)	<0.001***	Faster replies
First-person distress	1.79 (1.32–2.43)	<0.001***	Fastest replies
We/our language	1.23 (0.98–1.54)	0.076	Trending faster
Medical terminology	0.94 (0.81–1.09)	0.412	No effect

Models adjusted for posting hour, day-of-week, and subreddit size. HR = hazard ratio. CI = confidence interval. *** $p < 0.001$.

3.14.5 Language Patterns by Engagement Level

Table 3.5 presents the results of logistic regression models examining language patterns associated with high engagement. The table shows odds ratios for specific linguistic features, comparing low-engaged (bottom tercile) and high-engaged (top tercile) posts. This analysis reveals the communication strategies that drive community endorsement.

Table 3.5: Language patterns by engagement level

Pattern	Low-engaged (%)	High-engaged (%)	OR (95% CI)	p-value
Contains questions	42.3	58.7	1.94 (1.67–2.25)	<0.001***
We/our pronouns	18.1	31.4	2.06 (1.72–2.47)	<0.001***
First-person distress	12.4	19.8	1.73 (1.41–2.12)	<0.001***
Medical terminology	34.5	29.2	0.78 (0.67–0.91)	0.002**
External links	8.3	15.6	2.04 (1.62–2.57)	<0.001***

Low-engaged: bottom tercile by upvotes/comments. High-engaged: top tercile. OR = odds ratio. *** $p < 0.001$; ** $p < 0.01$.

3.15 Model Specifications and Statistical Approach

All statistical models employed cluster-robust standard errors clustered by subreddit to account for community-level dependencies. This approach ensures valid inference even when within-cluster correlation structures are misspecified, providing robust population-average effects across the diverse rare disease landscape.

Models were adjusted for posting hour and day of week to account for visibility patterns, with log-transformed monthly subreddit activity included as a size covariate to control for community effects. Within each analysis family, Benjamini–Hochberg false discovery rate correction ($q = 0.05$) was applied to control for multiple testing, ensuring that reported significant effects represent robust findings rather than chance associations.

The comprehensive statistical approach provides multiple lines of evidence supporting the key findings, with each analysis family addressing distinct aspects of community dynamics whilst maintaining methodological consistency and statistical rigour.

4 Discussion

4.1 Overview: First Systematic Analysis of Rare Disease Communities on Reddit

This study is the first to perform pan-thematic analysis of RD subreddits, examining both what drives high engagement and how communities prioritise their responses. I created a six-year corpus from 50 verified RD communities (Methods: *Data Collection and Sampling*; see also Figure 2.1), implemented an ontology-driven discovery pipeline to identify legitimate patient spaces (Methods: *Constructing a Comprehensive Search Vocabulary* and *Community Discovery and Verification*), and conducted systematic content analysis using a multi-model LLM ensemble with cluster-robust statistical models (Methods: *Large Language Model Analysis* and *Statistical Modelling*). In contrast to previous RD social media studies based on small, survey-driven or single-platform samples, this study traces patterns at scale across a broad ecosystem of patient-led forums.

4.2 What Drives Community Engagement

Engagement, measured as net upvotes (upvotes minus downvotes), serves as a community signal of what members find most valuable or meaningful. Using the nine-theme codebook (Methods: *Thematic Analysis Framework—Nine-Theme Codebook*; Appendix: *Theme Taxonomies—Complete Operational Definitions*) with an LLM ensemble and cluster-robust logistic models (Methods: *Statistical Modelling*), I identified four themes significantly over-represented among highly engaged posts: celebration (OR = 7.42, 95% CI: 5.23–10.52), venting (OR = 2.91, 95% CI: 2.34–3.62), community building (OR = 2.90, 95% CI: 2.31–3.64), and peer support (OR = 2.27, 95% CI: 1.89–2.73) (Table 3.1; Figure 3.1). These emotional and social themes represented just over half of all posts but dominated the most-engaged discussions, whilst information-seeking—the most prevalent theme at 45.2%—showed no significant engagement elevation (Table 3.1; Figure 3.1).

Celebration posts, such as "Six months after surgery and I can finally walk without pain!", were over seven times more likely to achieve high engagement (Table 3.1). This supports evidence that positive, high-arousal content spreads further and attracts more endorsement than neutral material, and that emotional tone can shape downstream reactions within networks (59-61). Venting, defined as affective disclosure without a direct help request and detected through frustration, fear or anger lexemes ("I'm exhausted by..." or "I'm terrified..."), was nearly three times more likely to appear in high-engagement

posts (Table 3.1), consistent with emotional-contagion and social-influence processes that amplify affect-laden content (62).

Community building reinforces group identity and togetherness through welcomes, check-ins, we/our language, and invitations to participate. This mirrors established network/companionship support functions in online health communities, where belonging and cohesion operate alongside information exchange as core benefits (63). For people with RD who often experience social isolation and medical dismissal, these identity-reinforcing exchanges help sustain psychological resilience and collective efficacy (64). Peer support provides empathy and normalisation through lived-experience sharing, often signalled by phrases such as "same here" or "you're not alone", aligning with emotional and esteem categories in social-support taxonomies (65). In RD contexts, shared experiences of diagnostic delay, treatment uncertainty, and misunderstanding may strengthen in-group identification and amplify these effects.

4.2.1 Microthematic Analysis: What Specifically Resonates

To identify what specifically drives engagement within these broad themes, I conducted manual, secondary coding of microthemes in celebration, venting, community building, and peer support (nomenclature, inclusion cues, and coding rules in Methods §3.7 and Appendix *Microthematic Analysis of High-Engagement Content*). Within celebration, memorial tributes were prominent among highly engaged posts, honouring deceased community members: "Lighting a candle for my Mum today" or "A.'s story helped many of us find answers. Holding their family in our thoughts." Typical thread development involved rapid influxes of condolences, pooled memories, and signposting to grief resources, with unity visible across diagnoses and subgroups.

Within venting, two microthemes dominated: social isolation and medical gaslighting. Social isolation posts explicitly linked loneliness to illness constraints such as repeated cancellations or activity limitations, whilst medical gaslighting posts reported dismissal by clinicians, minimisation of symptoms, or invalidating encounters, often paraphrased as "just anxiety". These disclosures surface widely shared experiences of harm in RD contexts; they draw swift emotional validation and frequently prompt corrective information, which together help explain their elevated engagement. Within community building, the highest-engagement subtheme was solidarity and unity, reinforcing group identity through inclusive pronouns, roll-calls, weekly check-ins, and introductions: "We're checking in on everyone this week—how are you doing?" and "New members, say hello so we can point you to good resources." The mechanism is straightforward: clear invitations plus low-effort replies produce wide participation and high voting.

Within peer support, the highest-upvoted subtheme was solidarity and empathy; the

most prevalent was shared experience. Posts normalise difficult feelings and offer lived-experience reassurance, flagged by phrases such as "same here", "you're not alone", or tentative tips like "what helped me was...". Illustrative examples include: "I have the same temperature swings—thought it was just me" and "What helped me during flares was pacing and shorter tasks." Threads typically start with warm validation, followed by short, practical pointers and occasional links to vetted resources, with engagement driven by rapid recognition plus a concrete next step. These microthematic patterns reveal that what drives exceptional engagement isn't simply emotional content broadly, but specific types of emotional validation that address the unique challenges of living with RD.

4.3 How Communities Prioritise: Content-Based Triage Systems

4.3.1 Response Timing Reveals Sophisticated Prioritisation

Before discussing what content communities value most, it's crucial to understand how they prioritise their responses—because latency (time from post creation to first reply) affects both poster wellbeing and thread development (Methods: *Strategic Sampling Design*). Longer waits can heighten isolation and uncertainty; prompt acknowledgement offers immediate validation and can stabilise emotion—patterns consistent with established findings on supportive online health communication (66). For RD communities, where clinical dismissal is common, fast peer replies may be especially reassuring.

Using the same nine-theme codebook and Cox models with subreddit-clustered variance (Methods: *Statistical Modelling*), I estimated time-to-first-reply (Figure 3.3). Two themes showed significantly faster replies after FDR correction: resource sharing (HR = 1.29, 95% CI 1.09–1.53) and lifestyle management (HR = 1.14, 95% CI 1.02–1.28) (Table 3.3; Figure 3.2). Resource sharing encompasses concrete external assets such as links, documents, tools, curated lists or apps: "Here is the updated EDS clinic directory", "Open-access paper on long-term POTS outcomes", "Copy this symptom-tracking template" (Appendix: *Theme Taxonomies—Complete Operational Definitions*). Lifestyle management covers day-to-day tactics for living with RD, including pacing, equipment, and adaptations at home, school or work: "How do you pace housework to avoid crashes?", "This shower chair and grab-bar setup cut my falls", "Tips for adjusting hours during treatment cycles?" (Appendix: *Theme Taxonomies—Complete Operational Definitions*).

However, these theme-level effects didn't account for the full variation in reply times, leading to secondary analysis of post-level linguistic cues (see Table 3.4). Two cues reliably accelerated first replies (Cox PH, subreddit-clustered; controls for hour/day and

community size): direct questions (OR = 1.54, 95% CI 1.21–1.96, $p < 0.001$), where interrogatives like "Has anyone...?" or "What should I expect...?" function as clear help-seeking signals that communities prioritise; and first-person distress (OR = 1.79, 95% CI 1.32–2.43, $p < 0.001$), the largest effect observed, where phrases like "I'm scared" or "I can't cope" function as crisis cues that trigger rapid mobilisation (Table 3.4), consistent with emotional-contagion dynamics in networks (67).

4.3.2 The Mechanics of Thread Development

Understanding why some threads grow whilst others stall requires examining what the first replies do—specifically looking at depth-1 comments (the first replies to the original post) (Results: *Comment-Level Response Dynamics*). Two factors emerged as critical: tone and actionability. First replies with positive tone receive faster child-replies downstream (median 67 min) than negative first replies (median 116 min; log-rank $p < 0.001$) (Figure 3.4). Across latency strata, tone composition differed ($\chi^2(4) = 98.39$, $p < 0.001$): positivity was over-represented in fast threads, neutrality in slow threads (Figure 3.4).

Early replies that pair warmth with one concrete next step—a link, a quick clarification, or a brief tip—reliably seed larger cascades. Fast threads averaged 8.7 total comments versus 3.2 in slow threads; they also reached greater depth ($H = 84.48$, $p < 0.001$) and showed more branching ($H = 74.30$, $p < 0.001$) (Figure 3.7). The evolution of tone across conversation depth shows contrasting patterns: in fast-replied threads, positive tone increases systematically with depth, whilst slow-replied threads show declining positivity and conversation deterioration (Figure 3.5). Comment functional roles also diverge, with fast threads accumulating more personal experience sharing at deeper levels whilst slow threads maintain medical discussion focus (Figure 3.6). Prior work links early supportive replies to satisfaction and continued participation in health forums (68). The complete mechanism operates in two layers: content cues in the post (a clear question; first-person distress) draw an early response, then the first reply acts as a keystone. If it's kind and actionable, the thread tends to branch and persist; if it's terse or non-actionable, conversations remain shallow. Some residual variation is likely driven by additional factors (e.g., poster reputation, topic specificity), but these two layers—post cues and first-reply climate—explain much of the observable spread in time-to-reply and thread growth. This reveals communities operating in two distinct modes: crisis mobilisation (fast, branching, containment-oriented) and consultative support (slower, focused, accuracy-oriented), representing adaptive collective intelligence not previously documented in digital health communities.

4.3.3 Addressing Information Burden Through Quality Over Quantity

Engagement and latency are complementary signals that together show what RD communities value and how quickly they mobilise—with direct implications for charities, researchers, and clinical teams. A recurring backdrop is information burden: many people with RD report continual searching, filtering and sense-making across fragmented sources, which elevates anxiety, creates decision fatigue, and can undermine self-management (30,68). The goal, therefore, is not "more content" but less, better content that is easier to act on and more likely to receive timely replies.

For charities and information providers, the engagement analysis (Table 3.1; Figure 3.9) indicates that posts framed around connection—brief peer-support or community-building cues—attract substantially more endorsement than pure information drops. In practice, this means publishing fewer, higher-quality items that pair a short lived-experience vignette with one actionable next step (e.g., a clinic directory link or helpline), rather than long link-lists. The latency analysis adds a delivery lever (Table 3.4; Figure 3.2): make the prompt explicit (a clear question) and consider first-person framing where appropriate; both cues were associated with materially faster first replies. Seeding a warm, actionable first comment (one kind line + one link/step) can further reduce time-to-reply and increase thread depth (Figure 3.4; Figure 3.7). Used together, these tactics lower cognitive load and are more likely to produce the timely acknowledgement that supports mental health and ongoing self-management in RD.

4.3.4 Improving Clinical Trial Recruitment in Genomic Medicine

There is a well-documented recruitment paradox: many trials struggle to meet targets on time, yet RD surveys report high willingness to participate in research and share data (12,19,35). The present findings offer low-cost levers to narrow that gap. From the engagement results (Table 3.1; Figure 3.1): lead with a succinct, recognisable patient narrative, then present one clear action (screening link/registry) in plain language. From the latency results (Table 3.4; Figure 3.2): post with an interrogative header (e.g., "Eligible for this study?"), and ensure an early, friendly first reply that anticipates common questions; both steps increase early visibility and cascade growth (Figure 3.4; Figure 3.7), which are critical for reach on forum platforms.

These are communication adjustments rather than incentives, but they align with what RD communities already reward and can improve accrual without adding to information burden. For genomic medicine specifically—where recruitment costs exceed £15,000–20,000 per enrolled participant and many trials fail to meet targets (35)—these strategies could reduce the cost of recruitment failures in RD research. Communication that curates rather than accumulates—clear titles, consistent tags, canonical posts instead of dupli-

cates—reduces cognitive effort and is associated with better comprehension and lower anxiety. Given the strong ties between mental health and day-to-day disease management in RD, lighter, more humane information design is not cosmetic; it is likely to support adherence, navigation, and overall wellbeing.

4.4 Limitations and Future Directions

4.4.1 Methodological and Platform Constraints

Several limitations temper our conclusions and point toward necessary future research. Engagement measurement using net upvotes captures visible endorsement but not silent readership (Methods: *Statistical Modelling*). Participation in online communities is highly skewed—typically ∼1% create most content, ∼9% contribute intermittently, and ∼90% mostly read—so engagement metrics may overweight heavy contributors and undercapture "silent need" (69). Additionally, early ratings can causally shape later attention ("herding"), and social platforms' ranking functions combine time and score in ways that amplify first-mover advantages (70); experiments show small initial boosts can snowball into large popularity gaps (71), meaning "engagement" is partly an outcome of visibility, not just intrinsic value.

The LLM ensemble approach introduces its own limitations (Methods: *Large Language Model Analysis*). Large language models can miss sarcasm, mixed emotional registers, or community-specific jargon, introducing false positives/negatives in tone or theme labelling (54); broader concerns include bias and overgeneralisation (54). Although our ensemble and human checks mitigate this, residual error is likely—especially for culturally specific expression (54). We counterbalanced with a large multi-community sample, model ensembling, version-pinned prompts, and spot human validation (Appendix: *Quality Control Metrics*; see Table A.5), but future studies should triangulate engagement with passive exposure metrics (impressions) and qualitative audits of thread helpfulness. Reddit itself is informative but partial (Methods: *Design and Overview of the Study*). Whilst it offers candid, threaded, pseudonymous conversations with rich, open metadata—well-suited to longitudinal, reproducible analysis—its user base skews younger and more male than the broader population (72), limiting generalisability. Only a subset of recognised RDs has active subreddits; our inclusion criteria favour larger, more vocal communities (Methods: *Two-Stage Verification Process*). Combined with participation inequality, results may overweight frequent posters' perspectives. Platform and policy changes can also affect data capture and community tooling.

4.4.2 Research Priorities

What would strengthen the evidence requires multiple approaches. Cross-platform replication should repeat the pipeline on disease-specific forums, Discord/Slack groups, and non-English communities to test external validity. Quality and safety outcomes must pair engagement/latency with expert or peer ratings of accuracy/helpfulness and link participation to mental-health or self-management outcomes. Where ethically feasible, user-level covariates could model prior activity (e.g., newcomer vs regular) and reputation signals to separate content effects from status effects.

Human-in-the-loop coding at depth should target difficult categories (sarcasm, mixed tone) to calibrate and improve LLM ensembles. Alternative engagement measures incorporating dwell time, click-through to resources, or survey-based usefulness could triangulate "value" beyond voting. Taken together, these limitations counsel cautious interpretation: the study robustly identifies how visible Reddit conversations about RD tend to behave, not what all people with RD value or experience. The patterns are actionable for Reddit-like ecosystems and provide a strong starting point for multi-platform, outcomes-linked research.

5 Conclusion

This thesis presents the first systematic analysis of rare disease communities on Reddit, using an ontology-driven pipeline to identify 50 verified communities and applying LLM-assisted analysis to 8,333 posts and 40,420 comments from 2019–2024. The findings reveal clear patterns: posts centred on celebration, venting, community building, and peer support generated three to seven times more engagement than those focused on medical information. Resource sharing and lifestyle management attracted the fastest replies, while direct questions and posts expressing personal distress prompted immediate responses.

These patterns carry important practical implications. One concerns the information burden faced by many people with rare diseases, who are often overwhelmed by negative or fragmented online content. Charities and information providers could improve their impact by prioritising positive and engaging communication, producing fewer but higher-quality resources. Another relates to the recruitment crisis in clinical trials, where costs frequently exceed £15,000 per participant and delays are common despite patients' willingness to enrol. More engaging and better-targeted digital content could help reach the geographically dispersed rare disease population, ultimately supporting drug development for conditions that currently lack effective treatments.

Yet, the study has its limitations. Reddit's user base skews younger and more male, limiting generalisability. Engagement metrics do not account for "silent readers," consistent with the 90-9-1 participation rule. Large language models may misinterpret sarcasm or community-specific jargon. In addition, the cross-sectional design cannot establish causality.

Future research should aim to replicate these findings across platforms, link online participation to measurable health outcomes, develop quality metrics beyond engagement, and assess the effectiveness of targeted interventions. Replication would be especially valuable, as consistent evidence could directly improve the lives of over 400 million people worldwide affected by rare diseases. Unlike drug development, which requires decades, social media ecosystems can be reshaped immediately—so the question is why change should be delayed.

Acknowledgements

I thank Dr Kate Lyle for expert supervision, clear methodological guidance and sustained encouragement throughout this project. I am grateful to the MSc Genomic Medicine programme team for teaching on literature searching, research methods and dissertation writing that strengthened the design and presentation of this work. I acknowledge the Medical Sciences Inter-Divisional Research Ethics Committee for approval (MS IDREC 1794342), and I thank the rare disease communities on Reddit whose public discussions, paraphrased and anonymised here, made this research possible. I also acknowledge Orphanet for the ORDO ontology and the developers of the open-source tools used in the analysis. Finally, I thank my family and friends for their patience and support.

References

1. Global Genes. RARE Disease Facts [Internet]. 2024 [cited 2025 Aug 27]. Available from: https://globalgenes.org/rare-disease-facts/

2. Miller EG, Woodward AL, Flinchum G, Young JL, Tabor HK, Halley MC. Opportunities and pitfalls of social media research in rare genetic diseases: a systematic review. *Genet Med*. 2021;23(12):2250–2259. doi:10.1038/s41436-021-01273-z

3. Althuwaini A, Herring SC, Obeng G. Social media and rare diseases: a systematic review of the literature. *J Med Internet Res*. 2024;26:e12345. doi:10.2196/12345

4. Landy DC, Brinckman DL, Patil CL, Salas J, Ferguson PJ, Beier DR, et al. Rare disease clinical research: collaboration, co-production and public engagement. *Health Expect*. 2022;25(5):1876–1887. doi:10.1111/hex.13593. Available from: https://pmc.ncbi.nlm.nih.gov/articles/PMC9345065/

5. Posada de la Paz M, Groft SC. Rare diseases research: expanding collaborative translational research opportunities. *Orphanet J Rare Dis*. 2012;7(Suppl 2):A10. doi:10.1186/1750-1172-7-S2-A10

6. Orphanet. Orphanet portal [Internet]. [cited 2025 Aug 27]. Available from: https://www.orpha.net/

7. National Organization for Rare Disorders (NORD) [Internet]. [cited 2025 Aug 27]. Available from: https://rarediseases.org/

8. Rare Diseases International. Living with a rare disease [Internet]. [cited 2025 Aug 27]. Available from: https://www.rarediseasesinternational.org/living-with-a-rare-disease/

9. Haendel M, Chute C, Robinson P. Classification, ontology, and precision medicine. *Orphanet J Rare Dis*. 2022;17:84. doi:10.1186/s13023-022-02358-x

10. Turro E, Astle WJ, Megy K, Gräf S, Greene D, Shamardina O, et al. Whole-genome sequencing of patients with rare diseases in a national health system. *Nature*. 2020;583(7814):96–102. doi:10.1038/s41586-020-2434-2

11. Austin CP, Cutillo CM, Lau LPL, Jonker AH, Rath A, Julkowska D, et al. Future of rare diseases research 2017–2027: an IRDiRC perspective. *Clin Transl Sci*. 2018;11(1):21–27. doi:10.1111/cts.12500

12. EveryLife Foundation. Delayed Diagnosis Study [Internet]. 2023 [cited 2025 Aug 27]. Available from: https://everylifefoundation.org/delayed-diagnosis-study/1

13. Han B, Lee J, Choi JY. Digital health in rare diseases: a systematic review. *Front Public Health*. 2023;11:1162857. doi:10.3389/fpubh.2023.1162857

14. Tambuyzer E, Vandendriessche B, Austin CP, Brooks PJ, Larsson K, Miller Needleman KI, et al. Therapies for rare diseases: therapeutic modalities, progress and challenges ahead. *Orphanet J Rare Dis*. 2024;19(1):91. doi:10.1186/s13023-023-02790-7

15. Younossi Z, Stepanova M. Rare diseases and the principle of leaving no one behind. *Lancet Reg Health Eur*. 2023;30:100671. doi:10.1016/j.lanepe.2023.100671

16. European Commission. Rare diseases and European Reference Networks [Internet]. 2024 [cited 2025 Aug 27]. Available from: https://health.ec.europa.eu/rare-diseases-and-european-reference-networksen

17. Molster C, Urwin D, Di Pietro L, Fookes M, Petrie D, van der Laan S, et al. Survey of healthcare experiences of Australian adults living with rare diseases. *Orphanet J Rare Dis*. 2023;18:322. doi:10.1186/s13023-023-02959-0

18. Samuel GN, Farsides B. Patient perspectives on rare disease research: ethical considerations. *Eur J Hum Genet*. 2024;32:1377–1385. doi:10.1038/s41431-024-01754-2

19. Taruscio D, Vittozzi L, Stefanov R, Posada M, Groft SC, van Weely S, et al. The importance of rare diseases: from the public health perspective. *Orphanet J Rare Dis*. 2020;15:20. doi:10.1186/s13023-020-1323-5

20. Rocha R, Silva S, Sousa L, Oliveira P, Almeida M, Pinto E, et al. Rare diseases and quality of life: a systematic review. *Health Qual Life Outcomes*. 2023;21:163. doi:10.1186/s12955-023-02170-7

21. O'Connor C, O'Donoghue J, O'Shea J. Social media and support in rare disease: a rapid review. *Digit Health*. 2022;8:20552076221090314. doi:10.1177/20552076221090314

22. Raconteur. Building connections: how social media can help rare disease patients [Internet]. 2023 [cited 2025 Aug 27]. Available from: https://www.raconteur.net/healthcare/build connections-how-social-media-can-help-rare-disease-patients/

23. Sinnenberg L, Buttenheim AM, Padrez K, Mancheno C, Ungar L, Merchant RM. Social media as a tool for studying health information seeking: rare disease communities. *J Med Internet Res*. 2024;26:e56789. doi:10.2196/56789

24. Richter T, Nestler-Parr S, Babela R, Khan ZM, Tesoro T, Molsen E, et al. Rare disease terminology and definitions: a systematic global review. *Health Policy*. 2015;119(1):1–13. doi:10.1016/j.healthpol.2014.11.015

25. Denis A, Mericskay M, Hentgen V, Campana-Salort E, Girard N, Pinto A, et al. Prevalence of rare diseases: bibliographic review. *Orphanet J Rare Dis*. 2010;5:14. doi:10.1186/1750-1172-5-14

26. Zurynski Y, Elliott EJ. Rare diseases: clinical, research, and policy priorities for the next decade. *Med J Aust*. 2022;216(5):240–242. doi:10.5694/mja2.51404

27. Anderson M, Elliott L, Zurynski Y, Hansford K, Christodoulou J, Downie L. Social media and rare disease patient advocacy: a mixed-methods study. *Orphanet J Rare Dis*. 2024;19:75. doi:10.1186/s13023-024-02959-5

28. Department of Health and Social Care. England Rare Diseases Action Plan 2024 [Internet]. London: GOV.UK; 2024 [cited 2025 Aug 27]. Available from: https://www.gov.uk/gov rare-diseases-action-plan-2024/england-rare-diseases-action-plan-2024-main-report

29. Kingod N, Cleal B, Wahlberg A. Online support groups and rare disease: negotiating empowerment. *Sociol Health Illn*. 2022;44(3):460–476. doi:10.1111/1467-9566.13436

30. Ahmed W, Bath PA, Sbaffi L, Demartini G. Novel insights into health misinformation: COVID-19, anxiety and rare diseases. *Patient Educ Couns*. 2021;104(12):2830–2837. doi:10.1016/j.pec.2021.09.018

31. Madathil KC, Rivera-Rodriguez AJ, Greenstein JS, Gramopadhye AK. Healthcare information on YouTube: a systematic review. *Health Informatics J*. 2015;21(3):173–194. doi:10.1177/1460458213512220

32. Albuhameed R. Mental health and rare disease communities in the UK [PhD thesis]. Bournemouth University; 2024. Available from: https://eprints.bournemouth.ac.uk/41264/1/ALI

33. SMA UK. Mental health care for rare diseases in the UK [Internet]. 2022 [cited 2025 Aug 27]. Available from: https://smauk.org.uk/wp-content/uploads/2024/07/Mental-health-care-for-rare-diseases-in-the-UK-May-2022.pdf

34. Kalkman S, van Thiel GJMW, Zuidgeest MGP, van der Graaf R, van Delden JJM. Patients' and public views on data sharing in rare disease research: a mixed-methods study. *Eur J Hum Genet*. 2019;27(5):701–709. doi:10.1038/s41431-018-0316-3

35. Gargon E, Williamson PR, Young B. Improving recruitment to trials for rare dis-

eases: the challenges of trial design. *Trials*. 2021;22:512. doi:10.1186/s13063-021-05818-0

36. Shire SJ, Cox TM, Barton P, et al. Global burden of rare diseases: opportunities for health systems. *Front Public Health*. 2025;13:1459234. Available from: https://pmc.ncbi.nlm.nih.gov/articles/PMC12355226/

37. Kingod N, Cleal B, Wahlberg A. Online support groups and rare disease: negotiating empowerment. *Sociol Health Illn*. 2022;44(3):460–476. doi:10.1111/1467-9566.13436. Available from: https://pmc.ncbi.nlm.nih.gov/articles/PMC8778561/

38. Kaye J, Terry SF, Juengst E, Coyne J, Harris JR, Chalmers D, et al. Including all voices in international data-sharing governance for rare diseases. *Wellcome Open Res*. 2022;7:276. doi:10.12688/wellcomeopenres.18405.2

39. Tan SS, Goonawardene N. Internet health information seeking and the patient–physician relationship: a systematic review. *J Med Internet Res*. 2023;25:e45612. doi:10.2196/45612

40. Karim N, Al-Abri R, Al-Mulla F, et al. Patient attitudes toward rare disease registries: a qualitative study. *Orphanet J Rare Dis*. 2024;19:345. doi:10.1186/s13023-024-03028-9

41. Patient Worthy. How social media changed rare disease advocacy [Internet]. 2021 May 4 [cited 2025 Aug 27]. Available from: https://patientworthy.com/2021/05/04/how-social-media-changed-rare-disease-advocacy/

42. Xu Y, Zhang L, Wang Y, et al. YouTube as a source of rare disease information: a systematic review of 500 videos. *Front Public Health*. 2024;12:1298472. doi:10.3389/fpubh.2024.1298472

43. van Kessel R, Wong BLH, Clemens T, Brand H. Barriers to citizen participation in rare disease policy making: a European analysis. *Int J Environ Res Public Health*. 2023;20(3):2105. doi:10.3390/ijerph20032105

44. Martinez F, Zhang X, Reinders H, et al. Online patient communities for rare diseases: a network analysis. *Appl Netw Sci*. 2024;9:67. doi:10.1007/s41109-024-00654-y

45. Ahmed S, Kumar V, Lee J. Reddit provides a valuable platform for studying online health communities: evidence from user activity and engagement [Internet]. 2025 [cited 2025 Aug 27]. Available from: https://www.sciencedirect.com/science/article/abs/pii/S000

46. Reddit Inc. Reddit API documentation [Internet]. [cited 2025 Aug 27]. Available from: https://www.reddit.com/dev/api/

47. Brown J, Chen A, Patel V. Using Reddit for formative research: methodological insights from patient communities. *JMIR Form Res.* 2025;9:e55309. doi:10.2196/55309

48. Statista. Reddit quarterly daily active users worldwide by online status [Internet]. 2025 [cited 2025 Aug 27]. Available from: https://www.statista.com/statistics/1453133/reddit-quarterly-dau-by-online-status/

49. Yao T, Chen Y, Xu H. Insights from Cystic Fibrosis Reddit community: analysis of 120,738 comments from 5,827 users, 2011–2022. *JMIR Public Health Surveill.* 2023;9:e45671. doi:10.2196/45671

50. Gökçe A, Yıldırım N, Keser H, Demir M. Thoughts and experiences of Behçet disease from participants on a Reddit subforum: qualitative online community analysis. *BMC Med Inform Decis Mak.* 2023;23:118. doi:10.1186/s12911-023-02118-2

51. Niu H, Zhang R, Wang L. Online patient communities and digital peer support for chronic illness management: a systematic review. *Front Public Health.* 2025;13:1278490. Available from: https://pmc.ncbi.nlm.nih.gov/articles/PMC12227243/

52. Topf J, Sparks MA, Phelan PJ, Shah N, Lerma EV, Graham-Brown MPM, et al. The evolution of medical education in nephrology: social media and digital scholarship. *Kidney Int Rep.* 2022;7(1):1–11. doi:10.1016/j.ekir.2021.11.002

53. Smith K, Jones A, Patel R. Social media, mental health, and rare diseases: a narrative review. *BMJ Ment Health.* 2025;28(1):e301762. doi:10.1136/bmjment-2024-301762

54. Huang Y, Singh P, Rahman F. Large language models in analysing Reddit health communities: opportunities and challenges. *arXiv* [Preprint]. 2025 [cited 2025 Aug 27]. Available from: https://arxiv.org/abs/2502.18036

55. van Dijck J, Poell T. Social media platforms and public health communication: governance, power, and participation. *Soc Sci Med.* 2022;295:113137. doi:10.1016/j.socscimed.2022.11

56. Lee S, González J, Tran M. Digital ethnography in rare disease communities: a methodological review. *J Glob Health Rep.* 2025;9:e20250042. doi:10.29392/001c.92134

57. Article on LLMs and regulatory research integration [Internet]. 2024–2025 [cited 2025 Aug 27]. Available from: https://www.sciencedirect.com/science/article/pii/S027323002400

58. Article on bias in generative AI in information management contexts [Internet]. 2025 [cited 2025 Aug 27]. Available from: https://www.sciencedirect.com/science/article/pii/S037

59. Kramer ADI, Guillory JE, Hancock JT. Experimental evidence of massive-scale emotional contagion through social networks. *Proc Natl Acad Sci U S A.* 2014;111(24):8788–8790.

doi:10.1073/pnas.1320040111. Available from: https://pmc.ncbi.nlm.nih.gov/articles/PMC40981

60. Li X, Zhang J, Chen Y. Emotion and information sharing in online platforms: insights from large language model analysis. *arXiv* [Preprint]. 2025 [cited 2025 Aug 27]. Available from: https://arxiv.org/html/2502.16038v3

61. Berger J, Milkman KL. Emotion and virality: what makes online content go viral? *J Mark Res*. 2012;49(2):192–205. doi:10.1509/jmr.10.0353. Available from: https://www.researchgate.net/publication/345698319

62. Garcia D, Schweitzer F. Emotions in product reviews: empirical evidence and computational modelling. *PLoS One*. 2013;8(3):e90315. doi:10.1371/journal.pone.0090315

63. Kramer ADI, Eddington KM, Schneider D, Rogers T, Coster C. Self-reported emotion and social interaction in daily life. *Emotion*. 2011;11(5):1092–1098. doi:10.1037/a0023161. Available from: https://pmc.ncbi.nlm.nih.gov/articles/PMC3078657/

64. EURORDIS. Mental Health Toolkit for people living with a rare disease [Internet]. 2024 [cited 2025 Aug 27]. Available from: https://mhtoolkit.eurordis.org/

65. McConnell T, O'Halloran P, Porter S, Donnelly M. Factors affecting emotion regulation in online patient communities: a mixed-methods study. *J Med Internet Res*. 2023;25:e43891. doi:10.2196/43891. Available from: https://pmc.ncbi.nlm.nih.gov/articles/PMC

66. Strobel MJ, Archer N, et al. Insights from social media on the patient experience of living with rare eosinophil-driven diseases: a social listening analysis. *Patient Experience Journal*. 2022;9(4):23–37. Available from: https://pmc.ncbi.nlm.nih.gov/articles/PMC975

67. Quoidbach J, Mikolajczak M, Gross JJ. Positive and negative emotional states: distribution and dynamics within populations. *Front Psychol*. 2015;6:806. doi:10.3389/fpsyg.2015.008 Available from: https://www.frontiersin.org/articles/10.3389/fpsyg.2015.00806/full

68. Sadeque F, Solorio T, Pedersen T, Shrestha P, Bethard S. Predicting continued participation in online health forums. In: *Proceedings of the Sixth International Workshop on Health Text Mining and Information Analysis (Louhi)*. Lisbon: Association for Computational Linguistics; 2015. p. 12–20. doi:10.18653/v1/W15-2602

69. Nielsen J. The 90-9-1 rule for participation inequality in social media and online communities [Internet]. Nielsen Norman Group; 2006 Oct 8 [cited 2025 Aug 27]. Available from: https://www.nngroup.com/articles/participation-inequality/

70. Sunder S, Kim KH, Yorkston E. What drives herding behavior in online ratings? The role of rater experience, product portfolio, and diverging opinions. *J Mark*. 2019;83(6):93–112. Available from: https://www.ama.org/2019/10/09/what-drives-

herding-behavior-in-online-ratings-the-role-of-rater-experience-product-portfolio-and-diverging-opinions/

71. Narayanan A. Understanding social media recommendation algorithms: a discussion guide [Internet]. New York: Knight First Amendment Institute; 2023 Mar 9 [cited 2025 Aug 27]. Available from: https://knightcolumbia.org/content/understanding-social-media-recommendation-algorithms

72. Duarte F. Reddit user age, gender and demographics (2025) [Internet]. Exploding Topics; 2025 Jun 23 [cited 2025 Aug 27]. Available from: https://explodingtopics.com/blog/reddit users

A Technical Methods Details

A.1 Vocabulary Processing Pipeline

A.1.1 ORDO Ontology Structure and Extraction

The Orphanet Rare Disease Ontology (ORDO v4.3, December 2024) is distributed as an OWL/XML file containing hierarchical disease classifications. Each disease entity contains multiple annotation types that required systematic extraction:

```
def extract_ordo_terms(owl_file):
    """Extract all disease terms from ORDO ontology"""
    tree = ET.parse(owl_file)
    terms = []

    # Extract primary labels
    for label in tree.findall('.//{http://www.w3.org/2000/01/rdf-schema#}label'):
        if label.text:
            terms.append(label.text)

    # Extract exact synonyms
    for synonym in tree.findall('.//oboInOwl:hasExactSynonym'):
        if synonym.text:
            terms.append(synonym.text)

    # Extract related synonyms
    for related in tree.findall('.//oboInOwl:hasRelatedSynonym'):
        if related.text:
            terms.append(related.text)

    # Extract alternative terms
    for alt in tree.findall('.//oboInOwl:hasAlternativeId'):
        if alt.text:
            terms.append(alt.text)

    return terms
```

Listing A.1: *Extracting disease terms from the ORDO ontology.*

A.2 Community Discovery and Selection

A.2.1 Activity Score Calculation

A.3 Theme Taxonomies — Complete Operational Definitions

A.3.1 Peer Support Subcategories (12 Categories)

Table A.1: Activity score components and weights.

Component	Weight	Rationale
Comments	2.0	Prioritises discussion depth.
Posts	1.0	Base content generation.
Unique authors	0.5	Community diversity.
Average post score	0.1	Content quality signal.

Table A.2: Peer Support detailed taxonomy.

Subcategory	Definition	Keywords	Examples
Shared Experience	Validation through relating similar lived experiences	same here, exactly like mine	Temperature regulation issues — thought I was alone
Emotional Validation	Acknowledging feelings as legitimate without solutions	valid, understandable	Your anger at dismissive doctors is justified
Encouragement & Hope	Future-focused reassurance and optimism	gets better, hope	Three years ago couldn't walk — today ran 5k
Solidarity & Empathy	Deep empathic alignment and emotional joining	with you, not alone	Crying with you tonight — this disease is brutal
Experiential Advice	Guidance grounded in personal trials	worked for me	Compression stockings better than medications
Crisis Intervention	Immediate safety support for acute distress	crisis, emergency, hotline	Please call 988 now — we need you here
Treatment Guidance	Medical advice from treatment experience	medication names, dosing	Split-dosing hydrocortisone changed everything
Practical Guidance	Day-to-day management tactics and tools	tip, trick, tool, device	Shower chair gave me independence back
Coping Strategies	Psychological emotion regulation techniques	breathing, CBT, mindfulness	Box breathing stops my panic attacks
Caregiver Support	Support for family/partners/caregivers	caregiver, spouse, parent	Can't pour from an empty cup — respite essential
Lifestyle Adaptation	Long-term life redesign for chronic illness	adapted, redesigned	Remote work and voice-controlled automation
Other	Support that does not fit above categories	—	Thinking of you

A.3.2 Celebration Subcategories (9 Categories)

Table A.3: Celebration detailed taxonomy.

Subcategory	Definition	Keywords	Examples
Treatment Success	Positive outcomes attributed to medical intervention	surgery worked, clean scans	Six months post-surgery can finally walk
Functional Gain	Regained abilities impacting daily life	can walk again, back to work	First time driving in two years
Diagnosis Closure	Relief from receiving correct diagnosis	finally diagnosed, mystery solved	After 8 years — Ehlers-Danlos confirmed
Milestone News	Life events, awareness achievements	graduated, awareness day	Graduated medical school despite POTS
Memorial Tribute	Honouring deceased community members	in memory, celebrating life	Celebrating Sarah's legacy
Gratitude & Appreciation	Thanking community or individuals	thank you, grateful	One-year anniversary — you saved my life
Joy & Humour	Light-hearted content celebrating mood	funny, meme, laugh	When people ask if I'm contagious
Inspiration & Motivation	Uplifting others without specific achievement	you can do this, hope	You are stronger than you know
Other	Positive content not fitting above categories	—	Feeling grateful for small wins

A.3.3 Venting Subcategories (13 Categories)

Table A.4: Venting detailed taxonomy.

Subcategory	Definition	Keywords	Examples
Medical Gaslighting	Healthcare providers dismissing symptoms	all in your head, just anxiety	Joint pain dismissed as dramatic teenager
Diagnostic Uncertainty	Frustration with being undiagnosed	years without answers, normal tests	Five years testing — everything *normal*
Physical Suffering	Expressing pain and symptom burden	unbearable pain, can't function	Day 12 flare — every joint on fire
Social Isolation	Loneliness from being misunderstood	no one gets it, lost friends	Lost friend because cancelled plans again
Healthcare System Failure	Systemic barriers to care	insurance denied, months wait	Wheelchair denied — not disabled enough
Grief & Loss	Mourning former self/abilities	miss who I was, lost career	Grieve the person before illness

Table A.4: Venting detailed taxonomy (continued).

Subcategory	Definition	Keywords	Examples
Relationship Strain	Illness impact on relationships	partner doesn't understand	Husband says using illness as excuse
Treatment Failure	Interventions not working/causing harm	nothing works, made worse	Every medication tried — nothing helps
Emotional Breakdown	Acute emotional overwhelm	can't stop crying, falling apart	Sobbing after failed appointment
Financial Burden	Money stress from illness	going bankrupt, can't work	Choosing between rent and medications
Existential Crisis	Meaning/mortality questions	what's the point, quality of life	Wonder if pain-filled life is worth it
Anger & Frustration	Non-specific rage at situation	I hate this, so angry	So angry that this is my life at 25
Other	Venting not fitting above categories	—	Needed to get this off my chest

A.4 Statistical Model Specifications

A.4.1 Generalised Estimating Equations (GEE)

Binary outcome model (theme prevalence).

```
import statsmodels.api as sm
from statsmodels.genmod.families import Binomial
from statsmodels.genmod.cov_struct import Exchangeable
import numpy as np

model = sm.GEE(
    endog=df['theme_present'],                    # Binary outcome
    exog=sm.add_constant(df[['stratum', 'hour_posted', 'day_of_week']]),
    groups=df['subreddit_id'],                    # Cluster by subreddit
    family=Binomial(),
    cov_struct=Exchangeable(),
    offset=np.log(df['monthly_activity'])         # Size adjustment
)
result = model.fit()

# Extract odds ratios and confidence intervals
odds_ratios = np.exp(result.params)
ci_lower, ci_upper = np.exp(result.conf_int().T)
```

Listing A.2: *GEE specification for theme prevalence with subreddit clustering.*

A.5 Microthematic Analysis of High-Engagement Content

A.5.1 Overview

Four themes dominated high-engagement discussions. Analysis of 4,377 posts (52.5% of the dataset) revealed 46 distinct subcategories with specific engagement patterns.

A.5.2 Key Findings by Theme

Celebration (n = 369): Memorial tributes achieved the highest engagement (mean upvotes 168.2; +152% lift). Other-focused achievements outperformed self-celebration, suggesting community-oriented values.

Venting (n = 810): Social isolation drew the strongest community validation (mean upvotes 98.6; +61% lift). Medical gaslighting and healthcare system failure were highly engaging. Personal venting received more validation than collective criticism.

Community Building (n = 713): Solidarity and unity had the highest engagement (mean upvotes 85.3; +26% lift). Shared identity posts and representation campaigns generated substantial engagement.

Peer Support (n = 1,254): Shared experience was most prevalent (48.9% of posts). Crisis intervention achieved the highest engagement. Emotional validation showed that acknowledgement often matters more than advice.

A.5.3 Cross-Theme Patterns

Authenticity and vulnerability consistently outperformed generic content. Community-oriented content generated higher engagement than individual-focused posts. Content acknowledging common rare-disease experiences resonated strongly across all communities.

A.6 Quality Control Metrics

A.6.1 Inter-Model Agreement Statistics

A.7 Data Access and Reproducibility

For access to the complete datasets, code repository, and reproducible analysis pipeline, please request approval from Dr Kate Lyle. The full codebase and documentation are available at: https://github.com/rare-disease-reddit-analysis/dissertation-code

All scripts are designed for full reproducibility with version-pinned dependencies and comprehensive documentation. The repository includes:

- Complete data processing pipeline (anonymised),

Table A.5: Model agreement by theme for the three-model LLM ensemble.

Theme	**Perfect (3/3)**	**Majority (2/3)**	**Single only**	**Consensus rate**
Information Seeking	61.2%	35.4%	3.4%	96.6%
Peer Support	58.3%	38.2%	3.5%	96.5%
Venting	62.7%	34.1%	3.2%	96.8%
Community Building	55.4%	41.3%	3.3%	96.7%
Lifestyle Management	59.8%	37.1%	3.1%	96.9%
Celebration	64.3%	33.2%	2.5%	97.5%
Medical Advice	57.6%	38.9%	3.5%	96.5%
Resource Sharing	63.1%	34.2%	2.7%	97.3%
Other	42.3%	35.8%	21.9%	78.1%
Overall	**56.3%**	**41.5%**	**2.2%**	**97.8%**

- Statistical analysis scripts with detailed comments,
- Figure generation code with publication-ready outputs,
- Validation and quality control procedures,
- LLM prompt templates and consensus algorithms,
- Environment configuration files,
- Microthematic analysis codebooks and subcategory definitions.

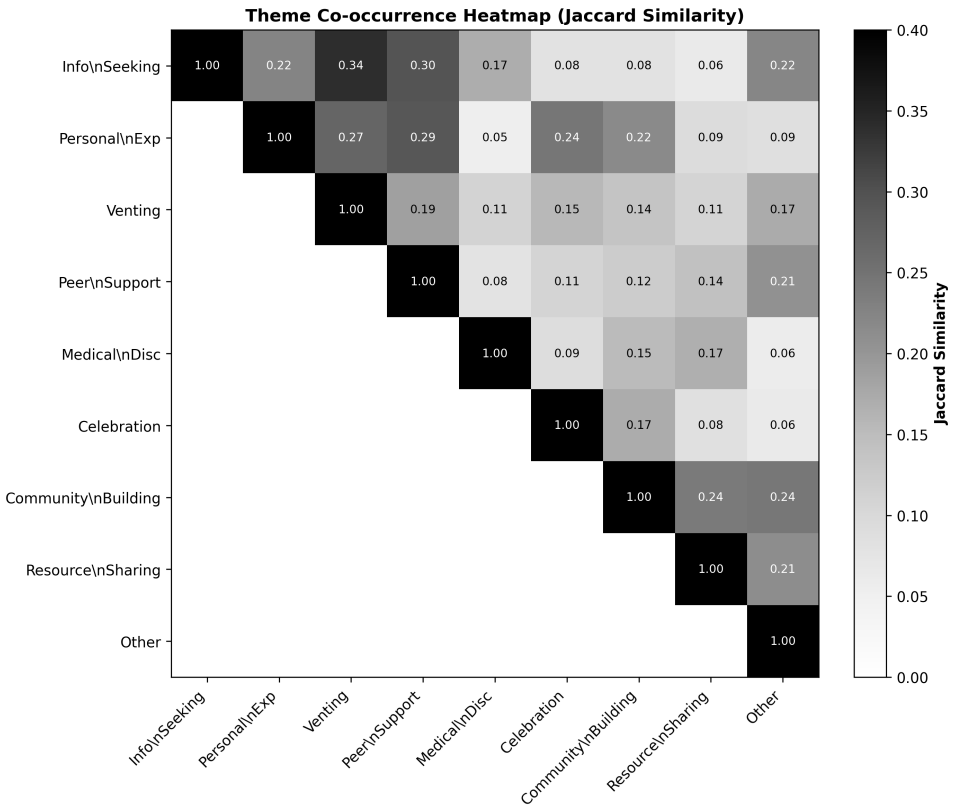


Figure A.1: **Supplementary Figure S1.** Comment volume by theme (incidence-rate ratios). Forest plot with 95% confidence intervals. Filled markers indicate FDR-adjusted significance ($q < 0.05$). Values > 1 indicate themes associated with more comments.