Aggregated Online Treatment Reports as Predictors of Clinical Trial Outcomes in Long COVID studies: A Methodology for Real-World Evidence at Scale

Polina Binder[1], Shaun Geer[1], Eli Sakov[1]

Abstract

Online patient communities for Long COVID contain thousands of self-reported treatment experiences, representing a naturalistic evidence base that has not been systematically analyzed. Here, we describe a pipeline to transform these posts into structured data by systematically extracting drug mentions, treatment sentiment, and other relevant information for analysis. We show that these aggregated treatment reports are consistent with published clinical trials and that social media posts from before a clinical trial has predictive value reflecting the outcome of the trial. These results suggest that systematically collected patient reports have predictive value for intervention efficacy.

Introduction

An estimated 10% of COVID patients develop Long COVID (Davis et al., 2023). There is no standard-of-care treatment, and there are no FDA-approved drugs for this condition. Because of this lack of clarity in the traditional biomedical literature, patients and clinicians are forced to rely on trial-and-error prescribing. Online patient communities have emerged to share treatment experiences, compare symptoms, and exchange practical guidance drawn from self-experimentation.

The scale of these communities is substantial. A descriptive summary of three of these communities about long COVID is below:

Table 1: Descriptive statistics of select long COVID patient community subreddits:

Subreddit	Posts	Comments	Total	Active Since
r/covidlonghaulers	117,351	1,926,869	2,044,220	July 2020
r/LongCovid	26,224	314,544	340,768	July 2020
r/LongCovidWarriors	877	9,313	10,190	June 2025
Combined	**144,452**	**2,250,726**	**2,395,178**	

Among these posts are informal treatment reports in which patients describe their attempted treatments and outcomes. These posts can be systematically extracted, structured, and aggregated.

[1] Patientpunk project, core contributors. https://github.com/Ely-S/PatientPunk . NCRI Health, Rutgers University. All authors contributed equally to this work. Contact email: team@patientpunk.org

Prior Work

Digital patient communities have increasingly been identified as a complementary source of Patient Experience Data within frameworks promoted by the U.S. Food and Drug Administration (FDA, 2020). Although treatment effectiveness and adverse event detection address different clinical questions, both tasks require recovering structured medical information from unstructured patient narratives and validating it against formal clinical evidence. The more extensively evaluated literature on adverse event detection can serve as a methodological precedent for the present study. And it has shown that social media can be used as leading evidence. Tricco et al. (2018) surveyed 70 studies and reports on social-media pharmacovigilance, of which 19 compared social-media-derived adverse events against external sources; 17 of those 19 reported broadly concordant findings, though the strength of comparison varied widely. Similarly, in a 2024 scoping review of 60 studies on social-media adverse-event detection, Golder et al. found that 78% (47/60) of these studies recommended social media as an adjunct to traditional pharmacovigilance.

Several recent studies have applied similar social media analysis pipelines to specific treatments. Ramagopalan et al. (2019) found that treatment patterns in renal cell carcinoma could be extracted from health-related social media forums using a combination of natural language processing, rule-based decisions, and machine learning. Kirchberger et al. (2025) found that Reddit narratives broadly corroborated Phase 3 RCT efficacy signals for ruxolitinib in vitiligo, and Sehgal et al. (2026) used over 400,000 Reddit posts to characterize the side-effect profiles of semaglutide and tirzepatide, confirming known gastrointestinal effects while surfacing unrecognized signals not captured in current labeling. In Long COVID research, large-scale NLP analyses have shown that platforms like Reddit can capture heterogeneous symptom patterns; for example, clustering over 27,000 posts identified distinct symptom groupings corresponding to clinically recognized Long COVID subtypes (Ayadi et al., 2023).

Together, these findings support the use of digital patient communities as a scalable, real-time complement to clinical data, particularly for characterizing heterogeneous conditions and generating hypotheses, while underscoring the need for careful validation. Prior quantitative work, however, has focused on treatment patterns (which drugs are used) or adverse events. Whether self-reported treatment effectiveness can be aggregated and validated against clinical trial primary endpoints has not been systematically tested, as this study attempts.

Limitations of Existing Clinical Long COVID Research

Long COVID intervention studies are plagued by small sample sizes, short follow-up periods, heterogeneous case definitions, poorly stratified patient populations, inconsistent endpoints, and limited blinding. Long COVID is multisystemic, fluctuating, and likely composed of multiple endotypes with different mechanisms and treatment responses. As Soares et al. note, “there is unlikely to be one outcome that can capture the scope of Long COVID’s progression or resolution,” and “single time-point assessments… fail to capture these dynamics” (2026). A recent meta-analysis likewise found that existing studies were limited by “small sample sizes” and “methodological heterogeneity” (Tan et al., 2025). As a result, many trials are underpowered to detect which subgroups benefit, and positive or negative findings may reflect endpoint choice, natural recovery, or patient mix rather than true treatment effect.

These weaknesses could be partly addressed through a large, standardized dataset of Long COVID case reports. At scale, structured real-world reports could identify symptom clusters, treatment-response patterns, adverse events, and clinically meaningful subgroups that individually underpowered statistical trials miss. Where clinical trials must commit to a single endpoint and a fixed population, community data can capture a broader subset of treatments and outcomes across a large number of heterogeneous subgroups that clinical trials struggle to stratify, even in meta-analysis. We do not view this approach as a replacement for traditional trials, but rather as providing a complementary signal that helps inform future clinical trials in which drugs, dosages, treatment durations, and subpopulations may be relevant to the study.

Methodology:

We set out to show if this large set of informal case reports can be exploited to advance the treatment development process. We hypothesized that pre-publication community sentiment toward a treatment would correspond to the eventual clinical trial outcome's direction; a binary framing that deliberately flattens trial heterogeneity into a test of whether the directional signal alone is recoverable from community discussion before publication. This would allow clinical researchers to enter into studies with more information about which interventions might be more effective. Clinical trials to compare social media posts against were chosen based on the following criteria:

- Trials that represented a diverse set of treatments
- Trials that looked at well-studied, commonly used drugs, which are well contextualized in the literature

We have chosen five studies in the literature and analyzed Reddit data that was posted before the study was published. These studies are shown in Table 2 and discussed below

Table #2 Comparative Clinical Trials

#	Paper	Drug(s)	Design	n	Outcome
1	Glynne et al. 2022	famotidine + loratadine (H1+H2)	Open-label combination	49	**Positive**
2	Utrero-Rico 2021 + Ye et. al. 2024	prednisone/corticosteroids	Small clinical study + meta-analysis	9 → 25 studies	**Null*** (The Utrero-Rico study was found to be positive, but not reproduced by larger studies)**
3	Geng et al. (STOP-PASC) 2024	paxlovid (nirmatrelvir-ritonavir)	RCT, double-blind	155	**Null**

4	Bassi et al. 2025	colchicine	RCT, double-blind	346	**Null**
5	O'Kelly et al. 2022	low-dose naltrexone (LDN)	Open-label pre-post	52	**Positive**

1)Glynne, P. et. al. (2022): *Long COVID following Mild SARS-CoV-2 Infection: Characteristic T Cell Alterations and Response to Antihistamines*
This observational study found that 72% of Long COVID patients experienced significant symptomatic relief after treatment with a combination of the H1 and H2 antihistamines, loratadine and famotidine. (Exact treatment varied but was typically loratadine 10 mg BID + famotidine 40 mg once daily for four weeks.) The study sampled 49 younger, predominantly female patients with Long COVID who had persistent symptoms following mild infections.
Outcome: positive.

2a) Utrero-Rico, A., et al. (2021): *A Short Corticosteroid Course Reduces Symptoms and Immunological Alterations Underlying Long-COVID*
Initial open-label pilot study of 9 Long COVID patients treated with a 4-day course of prednisone 30 mg/day. Deep immune profiling detected immunological alterations that resolved after the corticosteroid course, accompanied by symptom improvement.
Outcome: positive.

2b) Ye, X., Li, Y., Luo, F., et al. (2024): *Efficacy and safety of glucocorticoids in the treatment of COVID-19: a systematic review and meta-analysis of RCTs*
Sample: Meta-analysis of RCTs consisting of 5,721 hospitalized COVID-19 patients.
Methodology: Meta-analysis of randomized controlled trials evaluating corticosteroid use versus control, with subgroup analysis of specific steroids and disease severity.
Results: No clear overall mortality benefit in hospitalized COVID-19 patients, but corticosteroid class effects (including prednisone-like agents) showed improved outcomes in moderate-to-severe cases, without increased adverse events.
Outcome: null.

3) Geng, L. N., et al. (2024): *Nirmatrelvir-Ritonavir and Symptoms in Adults With Postacute Sequelae of SARS-CoV-2 Infection: The STOP-PASC Randomized Clinical Trial*
Double-blind, placebo-controlled RCT (n=155) of a 15-day course of nirmatrelvir-ritonavir (Paxlovid) for Long COVID at Stanford University. No significant benefit on any PROMIS-29 symptom severity endpoint in a mostly vaccinated cohort with protracted symptom duration.
Outcome: null.

4) Bassi, G. S., et al. (2025): *Effectiveness of Colchicine for the Treatment of Long COVID: A Randomized Clinical Trial*

Double-blind, placebo-controlled RCT (n=346) of colchicine for Long COVID across 8 hospitals in India. No improvement in functional capacity (6-minute walk test, p=0.45), respiratory function, or inflammatory markers at 52 weeks. The largest and best-powered RCT of colchicine for Long COVID to date.
Outcome: null.

5) O'Kelly, B., et.al. (2022): *Safety and efficacy of low dose naltrexone in a long covid cohort; an interventional pre-post study*
Open-label pre-post study of low-dose naltrexone (LDN, 1–4.5 mg/day) in 52 Long COVID patients. Improvement was observed in 6 of 7 quality-of-life components, with the largest effect in pain. Directionally replicated by Bonilla 2023 (n=59) and Isman 2024 (n=36); no RCT exists yet.
Outcome: positive.

Data and Sampling

Data was obtained from Arctic Shift (https://github.com/ArthurHeitmann/arctic_shift), a community project to back up and preserve Reddit data. We used the download tool to collect data from the following windows in the r/covidlonghaulers subreddit:

Table 3: Drug sampling windows and sizes

Drug	Window	Total posts	Unique users reporting treatment of interest	Treatment reports	Pre-publication cutoff
famotidine	2020-07-24 to 2021-06-05	138,615	231	692	Glynne et al.2022
loratadine	2020-07-24 to 2021-06-05	138,615	89	188	Glynne et al.2022
prednisone	2020-07-24 to 2021-10-25	218,822	343	790	Utrero-Rico et al. 2021
paxlovid	2020-07-24 to 2022-12-31	678,923	196	488	STOP-PASC Jun 2024
colchicine	2020-07-24 to 2022-12-31	678,923	91	211	Bassi et al. 2025
naltrexone	2020-07-24 to 2022-07-02	448,114	154	583	O'Kelly et al. Jul 2022

Our data includes both posts and comments in these timeframes. Famotidine, Loratadine, prednisone, naltrexone, and Paxlovid were sampled from the start of the subreddit to the day before the study was available to the public. Note that Glynne et al. 's findings first became publicly available as a medRxiv preprint on 2021-06-06. Colchicine and Paxlovid were sampled from the beginning of the subreddit until the end of 2022.

Filtering and data organization

Reddit organizes discussion as threaded reply chains: a top-level post may receive comments, which themselves receive replies. Many treatment discussions occur not in the original post, but in replies, where users respond to another user's mention of a drug with short statements like "same, it helped me too." To preserve this context, parent-child relationships between posts and comments were resolved at import. Our system allows the user to choose the length of the context chain; we found that preserving the two previous posts in the chain yielded the best results. Orphaned references, i.e. comments whose parent was not captured in the dataset, were cleaned to prevent broken chains. This linked structure is carried forward through the pipeline: during drug identification, each comment inherits the drugs mentioned in its parent chain, so that a reply discussing "it" can be correctly attributed to the drug named upthread.

Drug identification per post

Once the reply chains were pruned and established, our methodology then extracted Drug mentions using Claude Haiku 4.5, a fast, lightweight inference model suitable for high-volume structured extraction. For each post, the model returned all drugs, medications, supplements, and medical interventions mentioned, including brand names, generic names, abbreviations (e.g., "LDN"), drug categories (e.g., "antihistamines," "H1 blocker"), and informal references (e.g., "an oral antibiotic"). Diet and lifestyle changes were explicitly excluded, although they could be included in a broader exploration of this dataset. As mentioned before, the upchain drug context is important for this type of analysis: many comments were along the lines of “this worked really well for me,” and the information about what “it” is exists in a parent post upchain. Upstream drug context was computed by traversing the parent chain with caching, so that a reply three levels deep could still be linked to the drug introduced at the top of the thread.

Posts consisting entirely of questions (detected by sentence-final punctuation) were excluded from downstream classification, since they do not express personal treatment experience. However, these posts still contributed their drug mentions to the upstream context of any replies beneath them; a question asking "has anyone tried famotidine?" establishes the drug reference that a reply like "yes, it helped" depends on.

On batch failures, i.e. cases where the model returned a different number of results than posts submitted, the pipeline automatically split the batch into smaller sub-batches and retried, preventing silent data loss from truncated or malformed model output.

The LLM also sets a boolean flag if the user indicates they are using the intervention. Posts where the author is asking others about a treatment, citing research or studies, or offering encouragement without personal use (e.g., "Hope it works for you!") are filtered out. If the user does not indicate they are using the intervention, the post is excluded from the analysis to reduce noise in the data, but the information is still stored for context later in the reply chain.

Canonicalization / normalization

The extraction step can produce raw strings that should be grouped together: users refer to the same compound by brand name, generic name, or informal shorthand (e.g., “*Pepcid*,”"*famotidine*," and

"*famo*" should all be bucketed together in analysis). To aggregate reports at the compound level, we canonicalize mentions using a synonym-merging step.

The LLM receives the full list of extracted mention strings and identifies groups that refer to the exact same drug or compound. The key constraint is specificity: brand names and generic names for the same molecule are merged (e.g., "*pepcid*" → "*famotidine*"), and abbreviations are resolved (e.g., "*ldn*" → "*low dose naltrexone*"), but a specific drug is never collapsed into a broader pharmacological category (e.g., "*famotidine*" and "*antihistamines*" remain separate entries). The most common or recognizable name is selected as the canonical form.

The output is a mapping of non-canonical names to their canonical form; names with no synonyms in the input are treated as canonical by default. This mapping is applied to all downstream analyses so that per-drug aggregation reflects compounds rather than surface strings.

For per-drug analysis, a separate alias-lookup step generates a list of known names, brand names, abbreviations, and plausible misspellings for each target compound, ensuring that query-time filtering captures all relevant mentions.

Sentiment processing and Signal strength assignment

Posts that passed extraction were classified for sentiment using Claude Sonnet 4.6, a higher-capability model selected for this step because sentiment judgment requires interpreting nuance, hedging, and implicit meaning, which a faster model handles less reliably (Liang et al., 2023). Classification was grouped by drug; all posts mentioning a given drug were processed under a shared system prompt naming that drug and its known synonyms so that the model's frame of reference remained consistent within each drug's batch.

Each post received two labels. The first, sentiment captures the direction of the author's reported experience:

- **Positive**: The author personally used the treatment, and it helped.
- **Negative:** the author personally used the treatment, and it did not help or made things worse.
- **Mixed**: reserved for genuine ambiguity where symptoms actively worsened on one axis while improving on another (e.g., "helped my fatigue but made my anxiety worse"). Partial improvement, dose-titration struggles, and side effects during overall benefit were coded positive, not mixed.
- **Neutra**l: The author did not express a personal outcome or directional sentiment.

The second label received by the post, Signal strength, reflects the definitiveness of a statement that falls within each of these categories. This variable drives the deduplication tiebreak: when a user has multiple reports for the same drug, the report with the strongest sentiment and signal wins. This ensures that a user's most definitive statement, not their most ambiguous one, represents their experience in the data. Signal strength is labeled as **strong**, **moderate**, or **weak**.

Reply-chain context was available to the classifier: upstream comment text was provided, allowing the model to resolve pronoun references. However, the signal had to come from the reply itself; a reply expressing no personal experience was coded neutral even if the parent comment discussed the drug extensively.

Deduplication

A single user may post about the same drug multiple times. They may have an initial reaction, a follow-up months later, and then a reply in someone else's thread. To prevent prolific posters from dominating the signal, we deduplicate to a single report per user-drug pair.

When a user has multiple classified reports for the same compound, we retain their most recent report by full UTC timestamp. A user who posts a neutral dosage question in March and later reports a positive outcome in July contributes only the July report; conversely, a user who posts an enthusiastic positive in February and a follow-up "it stopped working" in May contributes only the May report. Signal strength (strong > moderate > weak) serves as a tiebreaker on exact-timestamp ties.

This yields a clean one-user-one-vote dataset for the between-group analysis: each user is either a responder or not for a given drug, with no double-counting.

Results

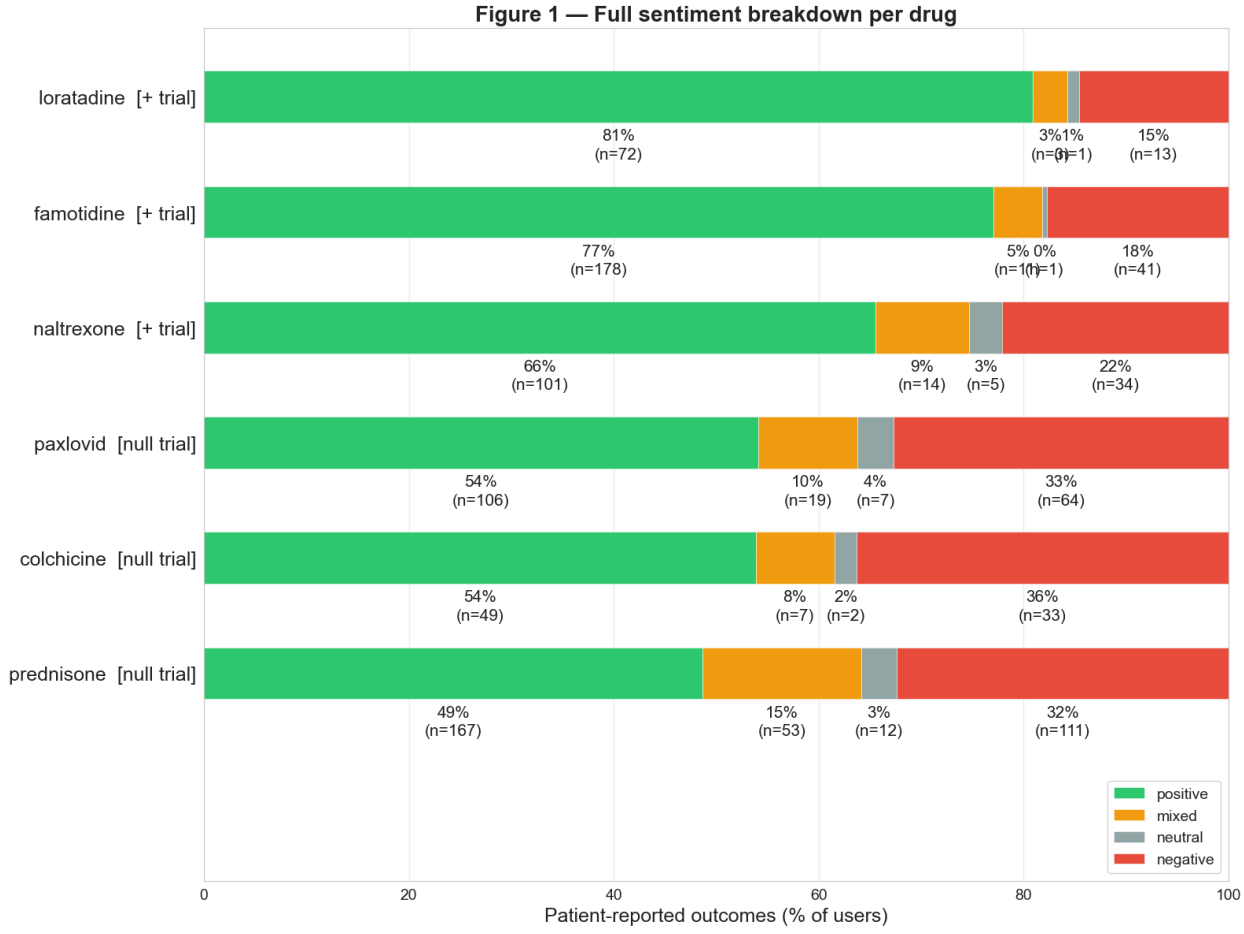

Figure 1

Figure 1 shows the four-way sentiment breakdown for each drug during the pre-publication window, compared with the corresponding clinical trial. Wilson score intervals rather than Wald intervals due to sample distribution asymmetry.

For drugs with positive trial outcomes (loratadine, famotidine, naltrexone), the positive bar dominates, and negative+mixed+neutral together stay small. For drugs with null trial outcomes, the picture differs: Paxlovid (54%/33% pos/neg, 10% mixed) and prednisone (49%/32% pos/neg, 15% mixed) have substantial mixed sentiment, suggesting users report partial responses or trade-offs. Colchicine has the most polarized null-trial distribution (54% pos / 36% neg, only 8% mixed). Note that the comparatively high mixed sentiment for prednisone suggests there may be partial or complex responses, which is different from a more negative result. If we collapse the non-positive sections into a "non-responder" category and compare responders to non-responders, we get Figure 2:

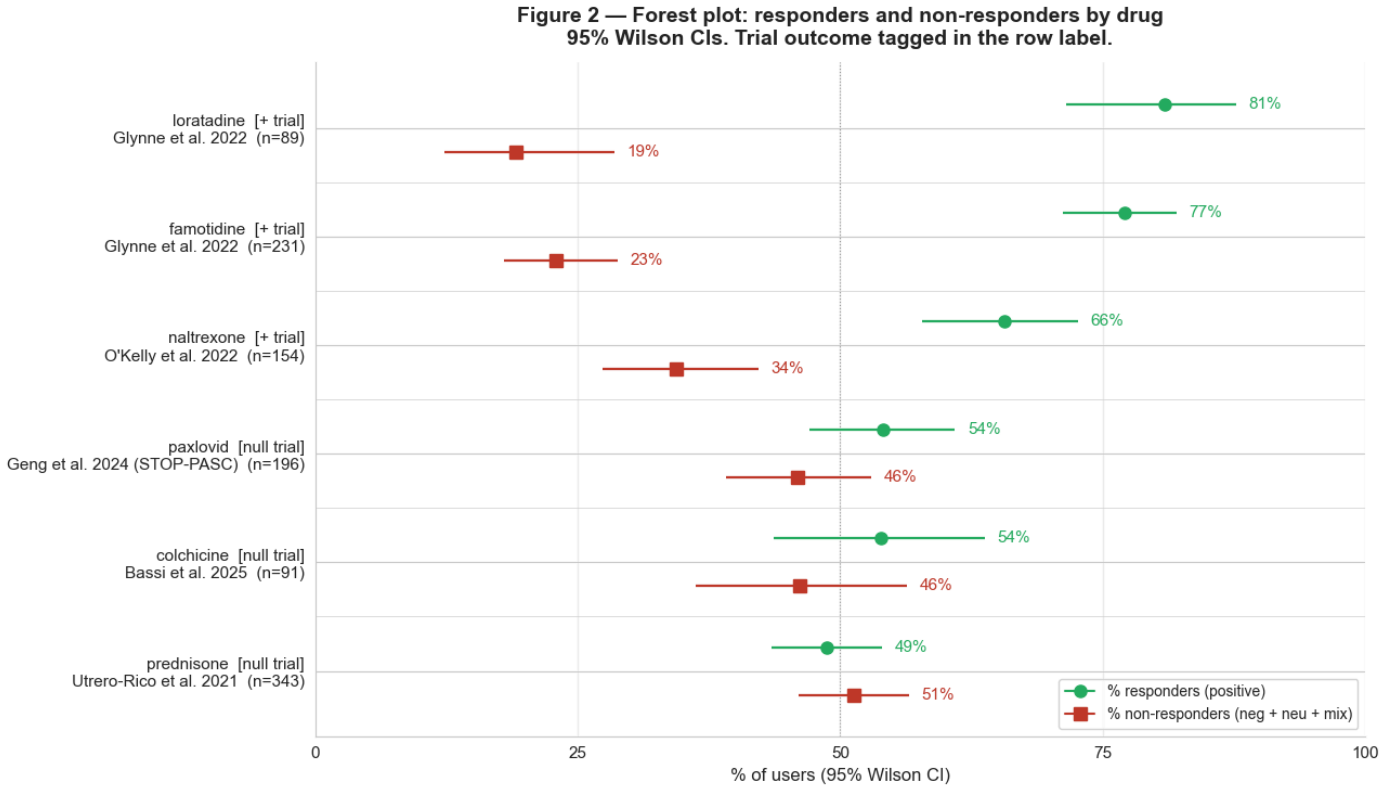

Figure 2

Compares confidence intervals (CIs) for the percentage of responders and non-responders from our analysis, with the outcome from clinical trials.. *Confidence intervals are 95% Wilson score intervals*. The right sidebar reports the clinical trial outcome that the Reddit data is being compared to. **For every** positive trial, our CIs estimated from online posts are mutually exclusive. For every null trial, our CI are overlapping.

In Figure 2, the 95% confidence intervals (CIs) for responders and non-responders differ significantly across all treatments with positive trial outcomes. For all treatments with negative trial outcomes, our CIs do not differ significantly, again matching the clinical trials. Under a null model in which community sentiment is unrelated to trial outcome, the probability of correctly classifying all six drugs into their realized positive/null buckets by chance alone is $0.5^6 \approx 1.6\%$.

drug	paper	n	% responders	responders 95% CI	% non-responder	non-responder 95% CI	p (vs 50%)
famotidine	Glynne 2022	231	77.2%	[71.3%, 82.1%]	22.8%	[17.9%, 28.7%]	3.56e-17
loratadine	Glynne 2022	89	81.1%	[71.8%, 87.9%]	18.9%	[12.1%, 28.2%]	1.95e-9
Low Dose Naltrexone	O'Kelly 2022	154	65.6%	[57.8%, 72.6%]	34.4%	[27.4%, 42.2%]	1.36e-4
colchicine	Bassi 2025	91	53.8%	[43.7%, 63.7%]	46.2%	[36.3%, 56.3%]	0.530
paxlovid	STOP-PASC 2024	196	54.1%	[47.1%, 60.9%]	45.9%	[39.1%, 52.9%]	0.284
prednisone	Utrero-Rico et al. 2021/Ye et al.2024	343	48.7%	[43.4%, 54.0%]	51.3%	[46.0%, 56.6%]	0.666

Table 4

Table 4 lists the percentage of responders to each treatment for each study. The P-value is the probability of obtaining the observed results under the null hypothesis that the responders and non-responders are drawn from the same population by chance.

For the three drugs paired with positive trials, we see positive sentiment significantly dominating our data. For the three drugs paired with null trials, we find statistically indistinguishable results from the literature ($p>.15$). These numbers depend on an LLM classification pipeline. Inter-rater reliability with two domain-trained human coders is reported in the appendix; the personal-use classification reaches a reliable Krippendorff's α (all pairwise $\alpha \geq 0.85$), while drug-extraction and sentiment α are in the trending-but-not-established range.

Discussion

The data analyzed above suggest that a simple between-group test on this data, with a relatively small sample size compared to the overall proportion of the subreddit, accurately predicts the results of all the studies listed above. Pre-publication community sentiment correctly classified the direction of outcome for all six drug–trial pairings. For every drug paired with a positive trial, those CIs do not overlap; for every drug paired with a null trial, they do. This pattern holds across drugs with very different mechanisms, trial designs, and patient populations.

Looking at community data in this way can provide additional signals that complement the conventional literature, at a lower cost than other methods for obtaining treatment data. An example of where this would have been useful is in the study of prednisone: our data suggested that the first 2021 study,

involving 9 patients, may not be generalizable to the population of long COVID sufferers. This data would have been available to both the community of Long COVID sufferers and medical professionals far sooner than any followup clinical studies.

One interpretation of these results is that r/covidlonghaulers functions as a large, slow, naturalistic trial running in parallel with the formal clinical research. Long COVID patients are unusually motivated self-experimenters: they track symptoms carefully, document changes over time, and update their reports as their condition evolves. The subreddit's sheer size also matters: even after restricting to personal-use reports and deduplication, per-drug sample sizes range from 89 to 343, comparable to the smaller trials in our comparison set. Where a 9-patient pilot can be dominated by chance variation, a community of hundreds of independent self-experimenters is more likely to reflect the population signal. This does not mean community data is free from bias; it is not, but it does suggest the signal is stable enough to survive those biases at scale.

The prednisone finding illustrates the practical value of this approach. Utrero-Rico et al. (2021) reported a positive outcome in 9 patients treated with a short course of prednisone, a result that was biologically plausible and generated reasonable clinical interest. Our community data, drawn from 343 unique users who all posted before that paper's publication, told a different story: 49% positive, statistically indistinguishable from the 50% null. Subsequent formal evidence has converged on the same null finding: the Ye et al. (2024) meta-analysis of 25 RCTs found no generalizable benefit for corticosteroids in this population. That community signal was available years before the meta-analysis, at no additional cost. Had it been available to trialists in 2021, it might have tempered enthusiasm for a positive result in 9 patients and directed investigative resources elsewhere.

There is no reason to believe that this approach is limited to Long Covid. Any condition in which patients self-experiment with treatments and discuss outcomes in online communities could, in principle, be analyzed in the same way. The necessary conditions are: a condition in which patients self-experiment with treatments and are motivated to discuss outcomes publicly; a patient community large enough to generate meaningful per-drug sample sizes; and a set of candidate drugs with enough community exposure to generate classifiable reports. Long COVID satisfies all three, but so do several other conditions that have outpaced the clinical literature. ME/CFS, fibromyalgia, and chronic Lyme disease all have large, treatment-active online communities and a sparse formal evidence base.

Limitations:
Our analysis is to be considered a pilot study and can be improved upon in many ways:

Lack of granularity in the analysis pipeline: Treating interventions as simply working or not working grossly simplifies patient outcomes. Long COVID sufferers have been shown to be very diverse in symptoms and responsiveness to treatments. Our methodology could be refined to better understand side effects, the symptoms and conditions that may make a treatment more effective, and the exact meaning of patients' reports that a treatment has had mixed results. Despite our current lack of granularity in measuring these and other data particulars, we find the strong significance in our study very encouraging.

Unclear and unfitted comparison point: A related assumption underlies the statistical test itself. We compare the responder proportion to a 50% null, treating the absence of effect as an even split between positive and non-positive reports. But we do not know what the baseline positive-sentiment rate is for this community independent of drug efficacy. If Long COVID patients are systematically inclined to report positive outcomes, perhaps because they continue using and posting about treatments that provide even modest relief, while silently discontinuing those that do not, the baseline could be above 50%, and our test would be conservative. If the reverse is true and frustration with a difficult condition produces a negativity bias, the baseline could be below 50%. Establishing this baseline empirically, for example, by identifying treatments with strong null evidence across many conditions and measuring their community sentiment rate, is a natural next step for validating the methodology.

Treatment Selection: Unlike in controlled trials, our community dataset lacks visibility into the screening processes that precede an intervention. Treatment accessibility also varies: patients can self-initiate over-the-counter medications like famotidine and loratadine, while prescription drugs require a physician's oversight.

Given our data's strong concordance with formal clinical trials, we hypothesize that treatment selection in this cohort was not random but guided by reasonable clinical judgment. For instance, if patients without underlying histamine issues were indiscriminately trying antihistamines, we would expect high rates of negative outcomes from the histamine intervention.

Non-random distribution of Reddit data. We expect selection bias among this subreddit's users: both in the general Reddit population and because people who have nothing to say or add may not post. We expect that positive and dramatic experiences may be overrepresented. Reddit tends to skew younger, more tech-savvy, and higher-class (Reddit, n.d.).

Pre-publication contamination. We analyze our data on journal-publication dates, but preprints, trial registrations, and conference presentations may have shaped community sentiment before our cutoff. This is most acute for STOP-PASC 2024 and Bassi 2025.

Polypharmacy confounding. Long COVID patients typically try multiple treatments simultaneously. Per-drug sentiment is contaminated by co-treatment effects. We downgrade signal strength when a drug is part of a stack in our analysis, but there is likely a more evidence-based way to handle this.

Dangers of historical fitting. The community data strictly predates each trial's publication. However, this historical concordance is a weaker claim than prediction: retrospective designs and constraints allow for both degrees of freedom and constrictions that a prospective study does not. In particular, we chose drugs and drug classes that we knew were well studied and likely had significant community discussion with strongly opinionated postings before these studies were published. This carries a risk of our overfitting to the past, with results that may vanish in confirmatory trials (Wang 2007). It is our view that concordance in retrospective trials is a necessary, but not sufficient, condition for predictive value.

Conclusions

We demonstrated that it's possible to derive meaningful signals from a massive dataset of informal treatment reports. We note that patient experiences align with clinical trials and suggest that they can be mined to guide future clinical trials.

Specifically in Long COVID, where there is a shortage of high-quality clinical trials and a large corpus of posts, we showed that social media data can guide future research. That is, before a clinical trial gets off the ground, we can estimate the treatment response from social media data.

Next Steps

This approach can be extended to help study treatment-response questions when the patient population is vocal about their treatment outcomes online and clinical trial data is sparse. Some examples are treatment peptide treatment response, where online communities have outpaced in-vivo studies, and complex chronic conditions like ME/CFS, Chronic Lyme, and Fibromyalgia.

A possible extension of this work is prospective rather than retrospective. Before initiating a trial on a candidate drug for a condition with an active online patient community, investigators could query pre-existing community data to estimate the direction of effect. A drug showing strongly positive sentiment enters the trial with more prior information than a drug at chance. A drug near the null enters with a caution flag. This does not replace the trial, but it changes the prior and may influence dosing choices, endpoint selection, and sample size justification. At the marginal cost of a database query and an LLM call, this represents a qualitatively different kind of information than is currently available at the trial-design stage.

This paper validates the approach. Next steps are to perform similar analyses on larger datasets, segment patient populations by demographics and symptomatology, and examine subgroup effects in treatment response.

The Long COVID patient community is extremely diverse, and each case has a unique set of symptoms. Practitioners have struggled to study this population as a whole due to the diversity of disease endotypes (citation needed). We hope that by studying a large enough number of treatment reports and by segmenting a massive population by symptom, we can start to answer questions like “For which type of patient is this treatment more likely to be effective?”

References

Ayadi, H., Bour, C., Fischer, A., Ghoniem, M., & Fagherazzi, G.(2023). The Long COVID experience from a patient's perspective: a clustering analysis of 27,216 Reddit posts. Frontiers in Public Health, 11, 1227807. https://doi.org/10.3389/fpubh.2023.1227807

Bassi, A., Devasenapathy, N., Thankachen, S. S., Ghosh, A., Rastogi, A., Khan, R., Bahuleyan, B., Gummidi, B., Basheer, A., Sreelal, T. P., Bangi, A., Shaikh, Y., Sahu, D., Rathore, V., Bhalla, A., Samita, S., Blessan, M., Dipu, T. S., Jain, M., ... Jha, V. (2025). Effectiveness of Colchicine for the Treatment of Long COVID: A Randomized Clinical Trial. JAMA Internal Medicine. 185(12), 1462–1470 https://doi.org/10.1001/jamainternmed.2025.5408

Davis, H.E., McCorkell, L., Vogel, J.M. & Topol, E.J. (2023). Long COVID: major findings, mechanisms and recommendations. Nature Reviews Microbiology, 21, 133–146. https://www.nature.com/articles/s41579-022-00846-2

Geng, L. N., Bonilla, H., Hedlin, H., Jacobson, K. B., Tian, L., Jagannathan, P., Yang, P. C., Subramanian, A. K., Liang, J. W., Shen, S., Deng, Y., Shaw, B. J., Botzheim, B., Desai, M., Pathak, D., Jazayeri, Y., Thai, D., O'Donnell, A., Mohaptra, S., ... Singh, U. (2024). Nirmatrelvir-Ritonavir and Symptoms in Adults With Postacute Sequelae of SARS-CoV-2 Infection: The STOP-PASC Randomized Clinical Trial. JAMA Internal Medicine, 184(9), 1024–1034. https://doi.org/10.1001/jamainternmed.2024.2007

Glynne P, Tahmasebi N, Gant V, Gupta R. Long COVID following mild SARS-CoV-2 infection: characteristic T cell alterations and response to antihistamines. J Investig Med. 2022 Jan;70(1):61-67. doi: 10.1136/jim-2021-002051. Epub 2021 Oct 5. PMID: 34611034; PMCID: PMC8494538. https://pubmed.ncbi.nlm.nih.gov/34611034/

Golder, S., O'Connor, K., Wang, Y., Klein, A., & Gonzalez Hernandez, G. (2024). The value of social media analysis for adverse events detection and pharmacovigilance: Scoping review. JMIR Public Health and Surveillance, 10, e59167. https://doi.org/10.2196/59167

Kirchberger, M. C., Berking, C., & Eisenried, A. (2025). Real-world use of topical ruxolitinib in vitiligo: Retrospective cross-sectional mixed-methods infodemiology study of the r/Vitiligo subreddit. Journal of Medical Internet Research, 27(1), e78247. https://doi.org/10.2196/78247

Liang, P., Bommasani, R., Lee, T., Tsipras, D., Soylu, D., Yasunaga, M., Zhang, Y., Narayanan, D., Wu, Y., Kumar, A., Newman, B., Yuan, B., Yan, B., Zhang, C., Cosgrove, C., Manning, C. D., Ré, C., Acosta-Navas, D., Hudson, D. A., ... Koreeda, Y. (2023). Holistic Evaluation of Language Models.
Transactions on Machine Learning Research. https://arxiv.org/abs/2211.09110

National Academies of Sciences, Engineering, and Medicine. (2024). Selected long-term health effects stemming from COVID-19 and functional implications. The National Academies Press. https://doi.org/10.17226/27756 / https://www.ncbi.nlm.nih.gov/books/NBK607399/

O'Kelly, B., Vidal, L., McHugh, T., Woo, J., Avramovic, G., & Lambert, J. S. (2022). Safety and efficacy of low-dose naltrexone in a long COVID cohort: an interventional pre-post study. Brain, Behavior, & Immunity – Health, 24, 100485. https://doi.org/10.1016/j.bbih.2022.100485

Ramagopalan, S. V., Malcolm, B., Merinopoulou, E., McDonald, L., & Cox, A. (2019). Automated extraction of treatment patterns from social media posts: An exploratory analysis in renal cell carcinoma. Future Oncology, 15(31), 3587–3596. https://doi.org/10.2217/fon-2019-0406

Reddit. (n.d.). Audience Insights. Reddit for Business. https://business.reddit.com/audience-insights

Sehgal, N. K. R., Tronieri, J. S., Ungar, L., & Guntuku, S. C. (2026). Self-reported side effects of semaglutide and tirzepatide in online communities. Nature Health. https://doi.org/10.1038/s44360-026-00108-y

Soares, L., Davis, H., Spier, E., Walker, T., Davenport, T., Putrino, D., Peluso, M., & Vogel, J. M. (2026). eBioMedicine, 123, 106083. https://doi.org/10.1016/j.ebiom.2025.106083

Tan, C., Meng, J., Dai, X., He, B., Liu, P., Wu, Y., Xiong, Y., Yin, H., Wang, S., & Gao, S. (2025). Effects of therapeutic interventions on long COVID: a meta-analysis of randomized controlled trials. eClinicalMedicine, 87, 103412. https://doi.org/10.1016/j.eclinm.2025.103412

Tricco, A. C., Zarin, W., Lillie, E., Jeblee, S., Warren, R., Khan, P. A., Robson, R., Pham, B., Hirst, G., & Straus, S. E.(2018). Utility of social media and crowd-intelligence data for pharmacovigilance: A scoping review. BMC Medical Informatics and Decision Making, 18, 38. https://doi.org/10.1186/s12911-018-0621-y

U.S. Food and Drug Administration, Center for Drug Evaluation and Research (CDER), Center for Biologics Evaluation and Research (CBER), & Center for Devices and Radiological Health (CDRH). (2020, June). Patient-focused drug development: Collecting comprehensive and representative input — Guidance for industry, Food and Drug Administration staff, and other stakeholders. Docket No. FDA-2018-D-1893. U.S. Department of Health and Human Services. https://www.fda.gov/regulatory-information/search-fda-guidance-documents/patient-focused-drug-development-collecting-comprehensive-and-representative-input

Utrero-Rico, A., Ruiz-Ruigómez, M., Laguna-Goya, R., Arrieta-Ortubay, E., Chivite-Lacaba, M., González-Cuadrado, C., Lalueza, A., Almendro-Vázquez, P., Serrano, A., Aguado, J. M., Lumbreras, C., & Paz-Artal, E. (2021). A Short Corticosteroid Course Reduces Symptoms and Immunological Alterations Underlying Long-COVID. Biomedicines, 9(11), 1540. https://doi.org/10.3390/biomedicines9111540

Wang, R., Lagakos, S. W., Ware, J. H., Hunter, D. J., & Drazen, J. M. (2007). Statistics in Medicine — Reporting of Subgroup Analyses in Clinical Trials. New England Journal of Medicine, 357(21), 2189–2194. https://doi.org/10.1056/NEJMsr077003

Ye, X., Li, Y., Luo, F., Xu, Z., Kasimu, K., Wang, J., Xu, P., Tan, C., Yi, H., & Luo, Y. (2024). Efficacy and safety of glucocorticoids in the treatment of COVID-19: a systematic review and meta-analysis of RCTs. Clinical and Experimental Medicine, 24, 157. https://doi.org/10.1007/s10238-024-01405-0

Appendix:

Inter-rater reliability of the LLM classification pipeline:

Most other studies examining the relationship between social media sentiment and clinical outcomes rely on human coders. To look at this in a preliminary fashion, we ran an inter-rater reliability (IRR) pilot on 100 samples from r/covidlonghaulers, drawn in a stratified random sample from a 30-day r/covidlonghaulers scrape starting on 2026-03-11. Strata was weighted 50% AI-flagged drug mentions / 30% medication-keyword matches / 20% neither to stress-test pipeline precision, recall, and specificity, respectively. Four coders independently labeled each sample using the same written codebook: two domain-trained human raters, plus the two Anthropic models the production pipeline relies on: Claude Haiku 4.5, a lightweight Anthropic model which performs the bulk pre-screen and substring-filtered drug detection, and Claude Sonnet 4.6, a mid-weight model which performs the sentiment classification and alias canonicalization. Stratification was also performed by haiku: Because the strata themselves were defined by Haiku's drug-extraction output, drug-extraction α involving Haiku is therefore upper-bounded by within-model consistency. However, sentiment, personal-use, and signal-strength α are unaffected in irr statistics involving Haiku, because the strata do not depend on those labels.

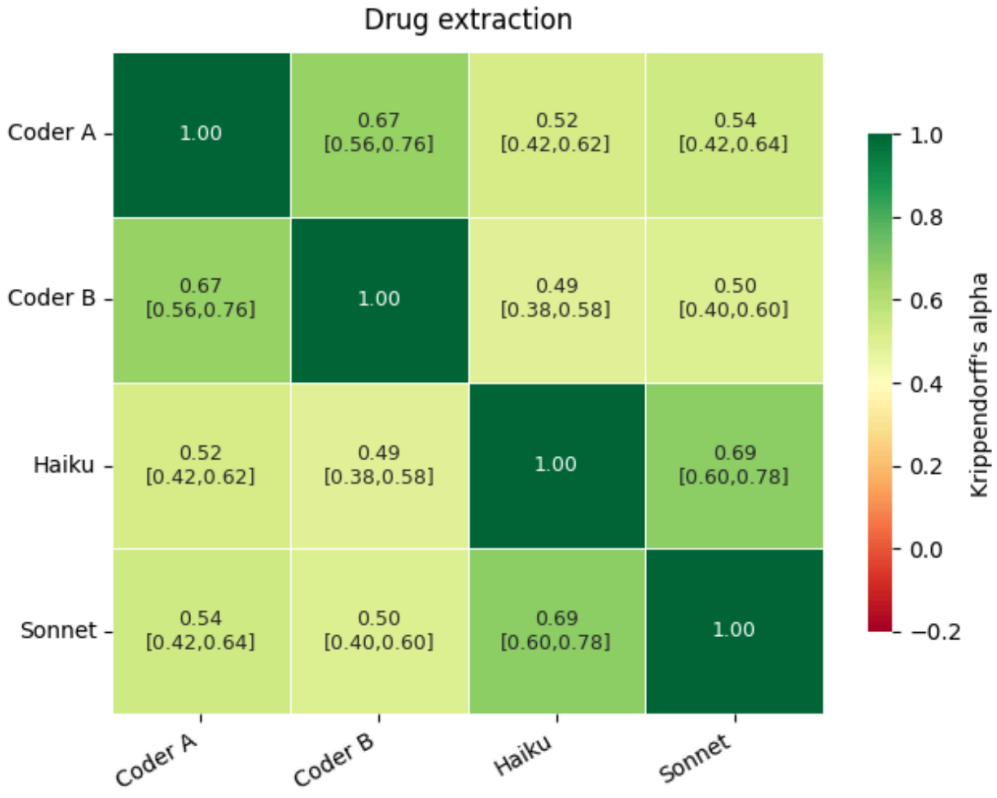

Figure 3
Pairwise inter-rater agreement on drug extraction. Krippendorff's α (nominal scale) across four coders

For every unit, each coder extracted any drug or treatment mention verbatim, then assigned, per mention, a (yes/no) personal-use flag, a four-way sentiment label (positive/negative/neutral / mixed), and a

signal-strength rating (strong/moderate/weak / n/a). Agreement was scored with Krippendorff's α on a nominal scale, the standard reliability metric for nominal categorical content coding because it corrects for chance agreement, accommodates missing labels, and is comparable across both the number of coders and the number of categories; conventional cutoffs are $\alpha \geq 0.80$ for reliable confirmatory use and $\alpha \geq 0.67$ for tentative exploratory use. The four-coder n-way α for drug extraction was 0.57 [95% bootstrap CI: 0.49 – 0.64], below the tentative threshold. Inspection of the disagreement traces this gap primarily to LLM over-extraction of non-drug interventions: fecal microbiota transplant, dietary changes, generic "supplements", which may indicate some noise being created by non-optimized prompting or that another model may more closely resemble human coding.

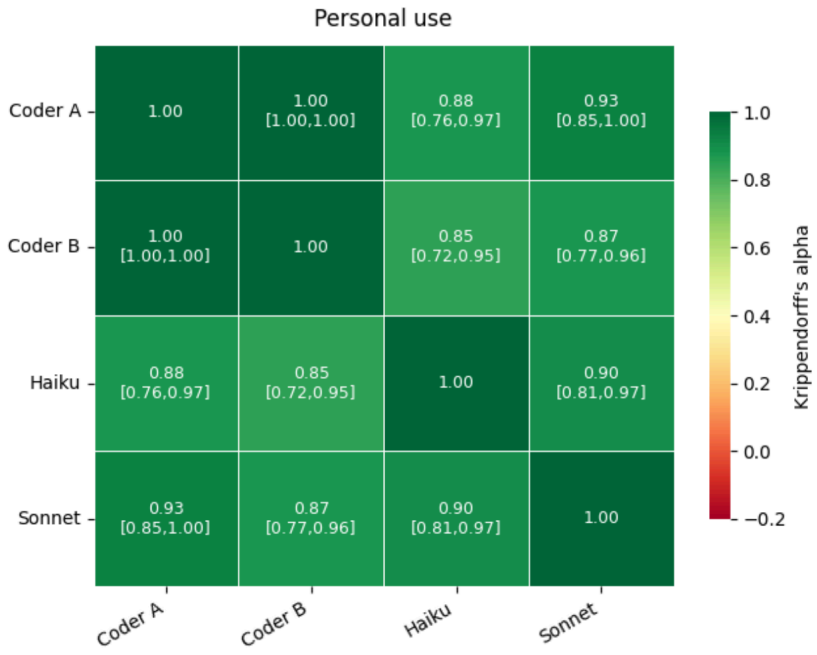

Figure 4

Pairwise inter-rater agreement on personal-use classification. Krippendorff's α (nominal scale) across four coders

On sentiment, conditional on a mention having been both extracted and marked as personal use, the n-way α across the four coders was 0.65 [0.57 – 0.72], again narrowly below the tentative threshold. The decomposition is informative: AI–AI sentiment agreement reached $\alpha \approx 0.81$, while human–human and human–AI pairs sat in the 0.60 – 0.70 range. The two LLM coders therefore apply a more internally consistent sentiment scale than the two human raters apply with each other, a pattern documented in instruction-tuned model evaluation, where rater calibration drift between humans is often larger than the within-model variance of a fixed checkpoint. Sentiment is the field that drives Figure 1 and Table 3 directly (responder = positive sentiment; non-responder = anything else), making it the most consequential reliability number for the paper's main claim. The present pilot value places it in the "trending" rather than "established" zone, distinguishable from chance, but not yet a textbook-reliable measurement.

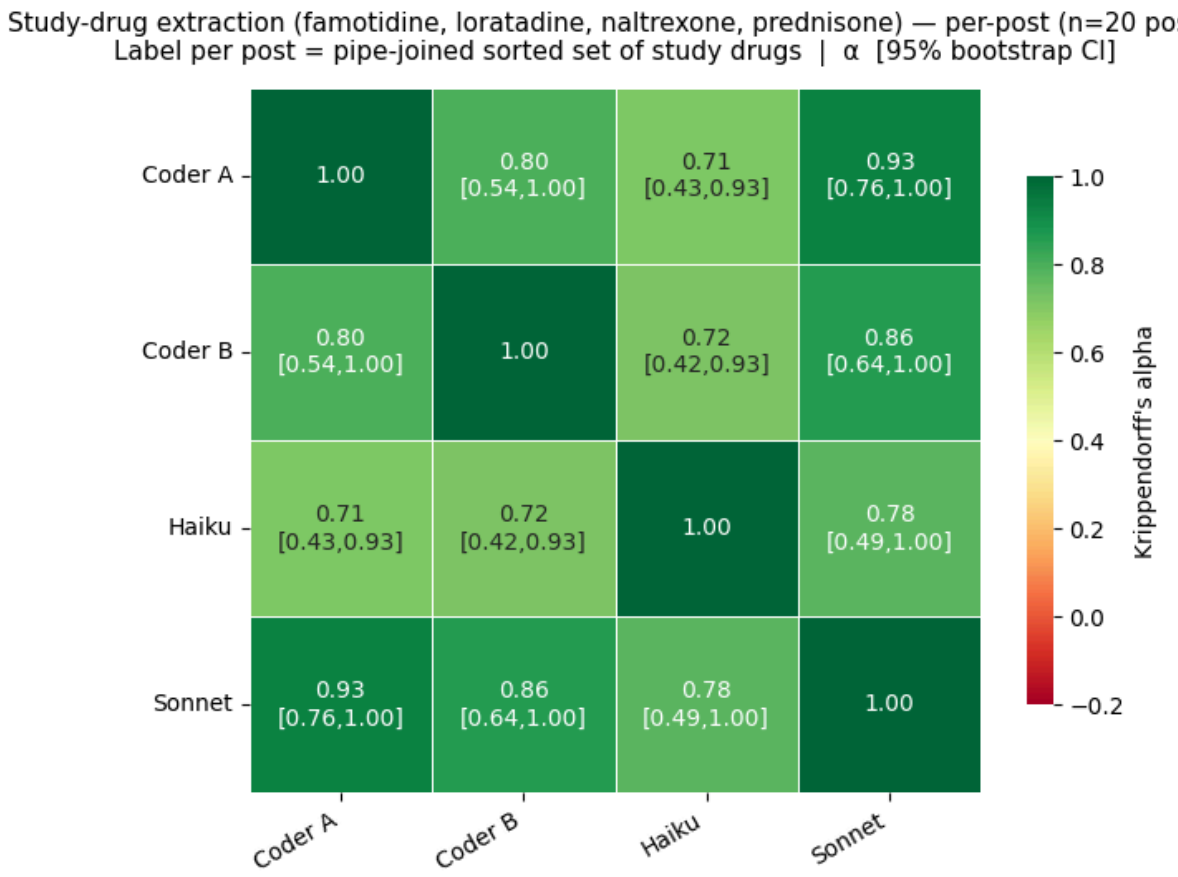

Figure 5

Pairwise inter-rater agreement on study-drug extraction, restricted to the four present in the IRR draw. Krippendorff's α (nominal scale) for the binary "did the coder extract this study drug?" question, pooled across famotidine, loratadine, naltrexone, and prednisone.

Because the paper's claim concerns the six pre-specified study drugs rather than the broader corpus of unrelated treatments, we also computed agreement specifically on the study-drug subset. Four of the six drugs (famotidine, loratadine, naltrexone, prednisone) appear in the 100-sample IRR draw; Paxlovid and colchicine are not present in the draw. Pooled at the post level i.e., counting every post in which at least one coder extracted any of the four present study drugs, yields 20 sample posts. On this subset, pairwise α for the binary "did the coder find this study drug?" question rises into the 0.63 – 0.94 range across coder pairs (median ≈ 0.75), and conditional sentiment α among study-drug mentions is correspondingly higher than the global figure. This is consistent with the over-extraction account: the noise in the global α comes from open-ended, non-study-drug content, not from disagreement about whether a mention of famotidine or low-dose naltrexone was a positive personal report. While this sample is too small to yield significance, it does trend to the LLM extractors and coders performing at a similar level to human coders for these four specific compounds that we discuss in our study. The system-level result reported in Figure 1, pre-publication community sentiment correctly ordering all six trial outcome, survives across coder, model, and dedup-rule sensitivity checks, so the open IRR status does not put the headline finding into serious question. Open methodological work remains: the expanded IRR run, per-coder calibration, model selection and prompt optimization, and a per-alias sensitivity ablation on the substring-match step.