Decide less, communicate more: On the construct validity of end-to-end fact-checking in medicine

Sebastian Joseph[1]* **Lily Chen**[2]* **Barry Wei**[3] **Michael Mackert**[1]
Iain J. Marshall[4] **Paul Pu Liang**[2] **Ramez Kouzy**[5] **Byron C. Wallace**[6] **Junyi Jessy Li**[1]
[1]The University of Texas at Austin, [2]Massachusetts Institute of Technology
[3]Indiana University School of Medicine, [4]King's College London
[5]The University of Texas MD Anderson Cancer Center, [6]Northeastern University
{sebaj, mackert, jessy}@utexas.edu, {l1ly, ppliang}@mit.edu, barrwei@iu.edu
iain.marshall@kcl.ac.uk, rkouzy@mdanderson.org, b.wallace@northeastern.edu

Abstract

Technological progress has led to concrete advancements in tasks that were regarded as challenging, such as automatic fact-checking. Interest in adopting these systems for public health and medicine has grown due to the high-stakes nature of medical decisions and challenges in critically appraising a vast and diverse medical literature. Evidence-based medicine connects to every individual, and yet the nature of it is highly technical, rendering the medical literacy of majority users inadequate to sufficiently navigate the domain. Such problems with medical communication ripens the ground for end-to-end fact-checking agents: check a claim against current medical literature and return with an evidence-backed verdict. And yet, such systems remain largely unused. To understand this, we present the first study examining how clinical experts verify real claims from social media by synthesizing medical evidence. In searching for this upper-bound, we reveal fundamental challenges in end-to-end fact-checking when applied to medicine: Difficulties connecting claims in the wild to scientific evidence in the form of clinical trials; ambiguities in underspecified claims mixed with mismatched intentions; and inherently subjective veracity labels. We argue that fact-checking should be approached and evaluated as an interactive communication problem, rather than an end-to-end process. Our data and code is available at `https://github.com/SebaJoe/decide-less-communicate-more`.

1 Introduction

Decision making in medicine is personal, intimate, and high-stakes. Traditionally the patient—often a lay person unfamiliar with medicine—converses with their care providers about questions about their health. However, the reality is far from this picture: most Americans resort to the web when they have a health-related question [Fox and Duggan, 2013]. Today, social media and AI have made medical knowledge seemingly accessible. But claims made by others on the web (or by an chatbot) can be inaccurate or inapplicable. This, combined with eroding health literacy [Champlin et al., 2017], have led to challenges in public health [Hassan et al., 2015] as well as patient-provider sessions.

Meanwhile, evidence-based medicine has continuously evolved, with an evidence base growing too rapidly for physicians to keep up [Bastian et al., 2010, Marshall et al., 2021]. This does provides a unique opportunity for AI fact-checking systems: Advances in retrieval systems and Large Language Models (LLMs) have increased interest in fact-checking systems that can classify medical claims as

*Equal Contribution

'True' or 'False' with supporting evidence. However, despite technological advances, such systems remain underutilized as they struggle to address diverse *claims in the wild*—naturally occurring statements, usually made by laypeople, that pervade public discourse [Das et al., 2023, Chen et al., 2022]; this further applies to claims made by AI agents that can be sometimes questionable. Existing fact-checking datasets often miss such claims because they were collected from already curated sources such as fact-checking websites and news articles [Kotonya and Toni, 2020, Vladika et al., 2024]. Prior datasets also extracted claims from their context [Sarrouti et al., 2021], generated synthetic claims [Saakyan et al., 2021], or filtered claims based on lexical criteria [Mohr et al., 2022]. As a result, systems trained on these heavily curated datasets are likely to fail to understand real medical claims made by the public.

To assess the boundary of AI-driven fact-checking systems, we focus on real-world medical claims from social media, preserving their original context. We contend that fact-checking systems should mirror how experts (e.g., physicians, care providers) evaluate and respond to such claims. We designed an annotation study that examines how medical experts verify claims using retrieved medical evidence. Experts were asked to assess medical claims present on social media, i.e., Reddit forums about a particular medical condition) by synthesizing retrieved randomized controlled trial (RCT) abstracts and explaining their judgments. This provides a novel benchmark for systems verifying *"in the wild"* claims. However, we highlight fundamental obstacles that challenge the construct validity of end-to-end automated systems for fact-checking: given a claim, provide a veracity judgment. We identify inherent difficulties in this setup for even domain experts, including: connecting claims with evidence; ambiguity from underspecified claims leading to valid yet contradictory interpretations; and challenges in achieving annotator consensus due to the inherent subjectivity of veracity labels. These issues suggest that the existing framing fact-checking of as an end-to-end classification task is inadequate for real-world settings, which may explain in part why such systems have not been put into wide use.

To correct the flawed construct validity of this task, we contend that **fact-checking should be an interactive dialogue agent rather than an end-to-end system**. We envision a human-centered **communication model** for medical fact-checking inspired by interactions between patients and physicians. We explain how this model can overcome existing challenges and empower experts and laypeople to engage in constructive medical discourse.

2 Background: medical claim-checking

Dataset	Domains	Source	Labels	Expl.	Evidence Type	Claim Example
PUBHEALTH (2020)	Public Health	Fact-Checking Websites, health news	True, Unproven, False, Mixture	✗	Sentences from same claim article source	Expired boxes of cake and pancake mix are dangerously toxic.
SCIFACT (2020)	Science	Expert-written	Supports, Refutes	✗	Scientific Articles	Rapamycin slows aging in fruit flies.
HEALTHVER (2021)	COVID-19	News Articles, blogs, social media	Supports, Refutes, Neutral	✗	Scientific Articles on COVID-19	Coronavirus may have originated in bats or pangolins
COVID-FACT (2021)	COVID-19	Reddit	Supported, Refuted	✗	Google Search Results	Baricitinib restrains the immune dysregulation in COVID-19 patients
COVERT (2022)	COVID-19	Twitter	Supports, Refutes, Not Enough Information	✗	Google Search Results	5G networks caused covid
REDHOT (2023)	Medical Conditions	Reddit	*N/A*	*N/A*	Randomized Controlled Trial Abstract	Link between RA and migraines
HEALTHFC (2024)	Health	Medizin Transparent	Supported, Refuted, Not Enough Information	✓	Systematic Review and Clinical Trial	Does cat's claw improve joint disease symptoms?
OUR CASE STUDY (2025)	Health	Reddit	No Relevant Abstracts, Refutes, Partially Refutes, Inconclusive, Partially Supports, Supports	✓	Randomized Controlled Trial Abstract	**Contextualized Claim** (3)

Table 2: A comparative view of related work in medical/health fact-checking. *N/A* indicates that the component is not applicable to the task (e.g., REDHOT does not perform claim verification).

Guo et al. [2022] outlines the conventional framework for automated fact-checking which comprises three stages: **(1) Claim Detection**, **(2) Evidence Retrieval**, and **(3) Claim Verification**. In the **Claim Detection** stage, the system identifies claims—statements asserting verifiable facts—and often ranks them based on check-worthiness factors such as public interest, popularity, timeliness, and impact [Das et al., 2023, Micallef et al., 2022]. Complex claims may also be automatically decomposed into sub-claims for individual verification [Wanner et al., 2024, Pan et al., 2023, Min et al., 2023, Kamoi et al., 2023a, Jing et al., 2024]. **Evidence Retrieval** entails retrieving supporting evidence to inform verification [Chen et al., 2024]. Finally, **Claim Verification** requires determining the claim's veracity and generating a justification grounded in the retrieved evidence. There is a growing interest in using LLMs to automate these stages of the fact-checking pipeline [Vykopal et al., 2024, Iqbal et al., 2024, Quelle and Bovet, 2024].

We present a comparative overview of prior work in medical fact-checking in Table 2. With few exceptions (e.g., Wadhwa et al. [2023] which we use in this study), existing work largely views claims as statements that "stand alone" without context, and has treated fact-checking as an end-to-end pipeline with the last step as a multi-label classification task [Sarrouti et al., 2021]. Additionally, no work has yet examined **expert involvement in every stage of medical claim checking** – a fine-grained examination of claim interpretation, retrieved evidence, and veracity judgment, with natural language explanations. As a result, we do not yet have an understanding of an upper-bound that modern systems could achieve on evidence-based medical claim verification. Finally, while formal evidence synthesis has long been studied in the health literature [Thoma and Eaves, 2015, Sackett et al., 1996, Moberg et al., 2018, Cumpston and Thomas, 2019], it has not been integrated into LLM-based medical fact-checking systems.

3 Case Study Methods

We formulate the following study for expert claim verification, in accordance to the task construction for automatic fact-checking systems discussed in section 2. We use claims from the Reddit Health Online Talk (RedHOT) corpus [Wadhwa et al., 2023], which contains 22,000 annotated posts from Reddit across 24 health conditions. RedHOT defines a claim as a statement indicating (often only implicitly) a causal relationship between an intervention and an outcome. To help experts contextualize these claims, we provide the full post and **P**opulation, **I**ntervention, **O**utcome (PIO) descriptors (illustrated in Table 3. See Appendix D,Appendix E for derivation details.) Note that we leave off the **C**omparators because in practice claims rarely mention the comparator explicitly (e.g., "Vitamin C cured my flu"). Given a contextualized medical claim, a medical expert constructs a hierarchy of evidence [Guyatt et al., 2008a,b, 2011] based on relevance and quality, assesses the claim's veracity, and provides a grounded explanation.

The overall task, without any aid, places a high cognitive load on experts who must comb through and synthesize multiple pieces of evidence [Juneja and Mitra, 2022]. With the aim of easing this, we automated several steps of the fact-checking process and integrated them into an intuitive web-based annotation interface (see Appendix Q). These features enable experts to focus on critical aspects: evaluating evidence relevance and synthesizing it to support or refute claims. Detailed annotation guidelines are provided in Appendix G.

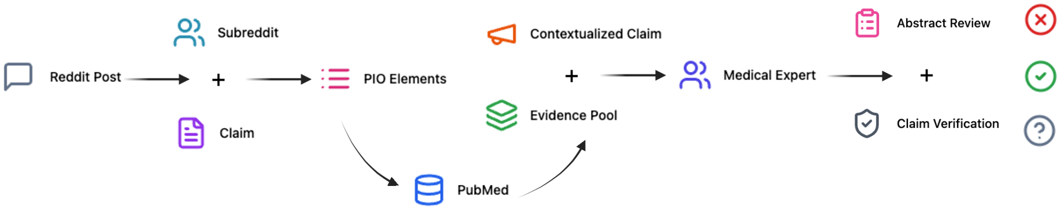


Figure 1: An overview of our AI-in-the-loop expert study pipeline: given a claim from a subreddit, we extract the PIO elements and retrieve the evidence automatically. The evidence, its context, and the evidence are then presented to a medical expert to provide a judgment and a rationale for the factuality of the claim.

For each claim, we used its PIO elements to automatically retrieve ten published relevant RCT abstracts from Trialstreamer [Marshall et al., 2020], a continuously updated database of RCTs, as potential evidence. We provide experts with these RCT abstracts and their publication dates. We used a dense retrieval system using state-of-the-art embedding models. We provided a more detailed description of this retrieval methodology in Appendix F. We do not evaluate the feasibility of manual evidence search in this study. However, experts do provide judgments on how relevant retrieved abstracts are to the claim in question. For each RCT abstract, we collected annotations determining its **relevance** to the claim along four dimensions: Population, Intervention, Outcome (PIO), and overall relevance. Claims are *contextualized* in their original Reddit posts during annotation. Annotators labeled each dimension as **(1)** Relevant, **(2)** Somewhat Relevant, or **(3)** Irrelevant. If an abstract was deemed *overall* relevant, annotators highlighted the most relevant text span and assessed whether the trial described in the abstract supports or refutes the claim using the four labels: **(1)** Supports, **(2)** Partially Supports, **(3)** Partially Refutes, and **(4)** Refutes.

After annotating all ten abstracts, experts proceeded to synthesize the evidence. To help experts navigate through the evidence documents, we implemented a tiering step. Abstracts are initially tiered automatically based on their relevance annotations in the previous step, establishing a natural hierarchy. Annotators are free to further refine this hierarchy by considering evidence quality. Next, annotators verify the claim in two phases:

1. **Overall Support:** Verification based solely on the provided evidence.
2. **Expert Support:** Optional verification incorporating their clinical experience and opinion. This is especially important for outlandish claims that are unlikely to have any trial evidence.

This separation allows us to compare expert opinion with evidence-based conclusions and minimize bias. Annotators select from six labels for both phases: **(1)** No Relevant Abstracts/No Expert Opinion (for each of the above phases respectively), **(2)** Refutes, **(3)** Partially Refutes, **(4)** Inconclusive, **(5)** Partially Supports, and **(6)** Supports.

To justify their veracity label, experts write a paragraph-length explanation (see guideline in Appendix M). Annotators may optionally include a medical addendum detailing clinical practices typically used in response to the claim, providing practical context for users.

Our annotation team consists of five clinical experts, one serving as the medical lead. All experts had experience reviewing medical articles and synthesizing them for biomedical research or patient care (annotator recruitment details are in Appendix H). To leverage their expertise effectively, we conducted the study in **three** rounds, with changes between rounds detailed in Appendix N. Following Klie et al. [2024], the two co-first authors held group meetings with the experts during each round to discuss disagreements and refine the annotation guidelines. In total, these meetings spanned four hours.

4 Results and Analysis

4.1 Challenge 1: Connecting Medical Evidence with Claims

We present the agreement on the final round of annotations for five claims annotated by five experts (totaling 25 separate verifications) in Table 5. This round includes 50 abstracts, with five labels per abstract, and two labels per claim. Despite multiple rounds of expert feedback to improve the annotation task, agreement remained low across all fields. A guideline for two annotators on two classes suggests that a Fleiss' κ score of 0.21-0.40 represents fair agreement, 0.41-0.60 moderate agreement, and 0.61-0.80 substantial agreement [Landis and Koch, 1977]. Since κ scores tend to be higher with fewer categories [Sim and Wright, 2005], we consider a reasonable κ score for our task with 4+ labels to be around 0.5.

Most instances—20 out of the 25 expert judgments—were judged to have "No Relevant Abstracts", indicating that the claims were unverifiable. For three of the five claims in this final round, all experts agreed they were unverifiable (see example in Table 4). This high rate of unverifiable claims underscores a broader challenge that "claims in the wild" are so often unverifiable given even the best conditions: The RedHot claims are annotated to "suggest (explicitly or implicitly) a causal relationship between an Intervention and an Outcome" [Wadhwa et al., 2023]; the evidence is retrieved with a state-of-the-art system.

Grapefruit

Post (r/Epilepsy): Grapefruit and siezures
Guys, I am wondering if you have any issues or know about interactions between oxcarbazepine and/or levetiracetam and grapefruit? I believe it may make those medications work differently, but I am not sure.

Population: Epilepsy patients (implied by the subreddit r/Epilepsy and the mention of seizure medications) ,

Intervention: Grapefruit consumption (in interaction with oxcarbazepine and/or levetiracetam) ,

Outcome: Medication efficacy (i.e., how the medications work)

Table 3: This claim is unverifiable because no RCTs have examined the interactions between grapefruit, Oxcarbazepine, and epilepsy, and conducting such a study may be infeasible.

Dandruff and Trikafta

Post (r/CysticFibrosis): Anyone had really bad flaky scalp or dandruff lately ? Think it could be due to trikafta. Could it be anything else

Population: Patients with Cystic Fibrosis (implied by the subreddit r/CysticFibrosis) ,

Intervention: Trikafta (a medication) ,

Outcome: Flaky scalp or dandruff

Expert 1: None of the abstracts directly addressed patients with cystic fibrosis experiencing dandruff/skin-related adverse effects secondary to trikafta use. All of the abstracts include some form of scalp flaking whether it be dandruff in general or specific conditions such as seborrheic dermatitis. However, they cannot be considered relevant as none of them address patients with cystic fibrosis or experiencing dandruff secondary to medication side effect.	**Expert 2:** A1, a2, a3, a4, a6 include the treatment of dandruff in a population with dandruff. The only relevant element in these abstracts is the outcome measures. A5, a7 includes a population with psoriasis and therefore, even the outcome measure here is irrelevant (i.e., all PIOs irrelevant). Similarly, a8 and a 9 included patients with seborrhoeic dermatitis, whereby the population and intervention were irrelevant, but scaling was an outcome measure (I would suggest somewhat relevant outcome). A10 also involved psoriasis population and intervention, but outcomes included scaling, therefore partially relevant outcome. Overall, the results are entirely inconclusive, since no abstract was relevant.	**Expert 3:** Overall, no relevant abstracts were available to analyze if trikafta may cause bad flaky scalp or dandruff. This is a very specific claim, and it is usually verified in the side effects results of RCTs. If not asked, participants may ignore the symptom if it is not significant.
Expert 4: There are no relevant abstracts to determine overall support. None of them include cystic fibrosis patients or the medication (Trikafta) from the claim. The outcome is mentioned in abstracts a1, a3, a4, a6, but has no relation to the claim.	**Expert 5:** The available studies included individuals with dandruff however none were diagnosed with cystic fibrosis, none were given Trikafta (a1, a2, a3, a4, a5, a6, a7, a8, a9, a10). Expert opinion: Dandruff can be caused by the underlying condition (cystic fibrosis) rather than as an effect of the medication itself (Trikafta)	

Table 4: All experts found the claim unverifiable based on the available RCTs. They attributed this to the claim's high specificity, noting it is unlikely—and potentially unethical—for a trial to match the described scenario.

Our expert annotation study identifies four reasons why:

1. **No Intervention:** Claims lacking an intervention cannot be verified through an RCT.
2. **Unethical Intervention:** Some interventions are unethical to test via RCTs because they may harm participants. For example, it is unethical to study smoking as an intervention in an RCT.
3. **Lack of Feasibility:** Claims involving specific PIO element combinations are often unverifiable due to the impracticality or improbability of conducting such RCTs.
4. **Lack of Utility:** Some claims, while theoretically verifiable through an RCT, lack available evidence as findings from such studies would lack utility in the medical field.

Type	κ (↑)
Population	0.416
Intervention	0.714
Outcome	0.200
Overall	0.155
Tab Support	0.170
Overall Support	0.124
Expert Support	-0.184

Table 5: Inter-annotator agreement for each portion of the fact-checking pipeline. Blue: abstract level labels, pink: synthesis level labels.

These issues highlight the difficulty of collecting evidence that is directly relevant to claims people make (on social media). Medical evidence, especially high-quality evidence, is bounded and restricted by standards for feasibility and ethics in a way that covers all the possible queries from the public is impossible. Prior to annotation, we used an automated method to filter out claims that could not

ADHD, Herbs, and Menstruation

Post (r/ADHD): Hello, menstruating people! How do your cycle and ADHD influence each other and how do you deal with it?
EDIT: After getting your responses I am reflecting again how medicine does not give a shit about women. It's truly insane. Thank you!
Hello! I have never paid too much attention to my menstrual cycle since it was never particularly bothersome. Now that I take methylo I feel big changes in how I function during the cycle. Like last 10 days of the cycle, my medication kind of stops working... That is like 1/3 of the time. I know it's still better than without meds nevertheless, it makes establishing a routine quite challenging. My doc suggested trying contraceptive pills, but I am not even sexually active ATM so taking more medication, with potential side effects, does not excite me. I know there are herbs that are proven to be helping with regulating the cycle but I don't know if they would help with ADHD symptoms? Any tips?

Population: People with ADHD, **Intervention:** Herbs, **Outcome:** Regulating the menstrual cycle

Expert Feedback:

- *What does regulating the cycle mean?*
- *ADHD has no bearing on one's menstruation cycle. It is a red herring.*
- *Trials with these descriptors are unlikely to exist.*
- *Is this claim really what the patient is concerned about in this post?*

Pineapple Juice Reduces Inflammation

Post (r/CysticFibrosis): Anyone with sinus issues drinking pineapple juice?
It's a weird question, but I saw a post about pineapple juice being good for sinus issues (helps with the inflammation) and just wondered if anyone has done this? Some people were commenting about the high sugar content in pineapple juice not being good, but they get around that by taking a supplement instead of drinking the juice. Anyone?

Population: Patients With Cystic Fibrosis, **Intervention:** Pineapple Juice, **Outcome:** Reduced Inflammation/Fewer Sinus Issues

Expert Feedback:

- *How quickly is the poster expecting the intervention to produce results?*
- *Just improving inflammation should not be the only criteria.*
- *Trials with these exact descriptors are unlikely to exist.*

Table 6: Examples of underspecified claims.

be verified by RCTs (detailed in Appendix K). This issue of unverifiability remained despite this effort. Part of solving the issue lies in addressing this issue of expanding the pool of evidence while ensuring it is still trustworthy (See Section 5.4). Additionally, systems should tackle how to best handle the inevitable case in which a claim is unverifiable. We discuss a guided retrieval approach to this in Section 5.2.

4.2 Challenge 2: Variations in the Interpretation of Claims

In our discussions with annotators, we noted consistent disagreements in how they interpreted medical claims, which in turn caused disagreement in the claim verification task. To address this, we provided annotators with PIO descriptors (Table 3) of claims to narrow the scope of possible interpretations. However, even with this added context reaching consensus among annotators remained a challenge. This is because *claims in the wild* about health tend to be *underspecified* and/or *misguided*, causing annotators to deduce their own varying interpretations on what the *patient*, the author of the claim (usually a layperson), intended.

Underspecified Claims Naturally occurring medical claims on social media are usually written informally by laypeople, and tend therefore to be underspecified (see Table 6). For example, a patient with ADHD claimed that "herbs are proven to be helping with regulating the cycle." It is unclear what "regulating" means here, and annotators interpreted this in various ways, e.g., reducing symptoms related to menstruation or skipping menstruation altogether. Another patient claimed that pineapple juice is "good for sinus issues" by reducing inflammation. However, it was unclear whether they meant immediate or gradual improvement (no time frame was offered). Resolving these underspecifications requires understanding the patients' intentions, which is challenging since intent cannot always be inferred from the claim and its context alone.

Misguided Claims Discussions with annotators also identified another artifact of naturally occurring medical claims: *misguided* claims formed from incorrect premises. Annotators often disagreed on how to handle such claims within our task. The previously mentioned ADHD example illustrates this issue. The patient, prescribed methylphenidate, noticed no effect during the last 10 days of their menstrual cycle. Their doctor suggested contraceptive pills as a potential solution. Concerned about the side effects of contraceptives, the patient considered using herbs as an alternative to "regulate the cycle". Annotators characterized the underlying premise—that the medication's efficacy is affected by the menstrual cycle—as incorrect. Disagreement over whether to consider the premise's validity in veracity judgments led to conflicting assessments among annotators. Prior work in general-domain fact-checking used claim decomposition to address false presuppositions in claims [Chen et al., 2022, Kamoi et al., 2023b, Hu et al., 2025], however this does not tackles *implicit* premises.

Mismatched Intent Previous work on fact-checking has addressed underspecified claims, often by decontextualizing them, i.e., removing context and resolving underspecifications based on local content [Deng et al., 2024, Gunjal and Durrett, 2024]. This approach can clarify underspecifications, but it disconnects the claim from the patient's original intent, as embedded in the global context. This disconnect can result in verifications that do not apply to the original claim, allowing subtle falsehoods to slip through and potentially be amplified. For example, suppose the underspecification of "regulating the cycle" were resolved and the claim were deemed true without considering its premise, this verification would fail to address the patient's true goal of improving the efficacy of their ADHD medication. Such verification would also implicitly validate the misunderstood premise. It is also the case that the patient's true intention is not about this particular claim, but an overall desire to communicate and discuss the underlying condition to get better. To effectively address this, the focus must be on meeting the patient's *information needs*, and when needed, uncovering and assessing their assumptions.

4.3 Challenge 3: Labeling the Severity of Inaccurate Statements is Inherently Subjective

Another factor that contributed to the disagreements we observed is the subjectivity in labeling the *veracity*, or the degree of truth of a medical claim. This subjectivity, as we observe, is not caused by differing interpretations of the claim. Rather, experts are influenced by their backgrounds and philosophies, applying different standards for assessing a claim's "truthfulness" based on evidence.

Consider the claim about pineapple juice in Table 6. If the desired onset for resolving sinus issues is clarified to mean within a few days, the claim is technically false. However, the *severity* of this falsehood is subjective. One expert might view it as a minor inaccuracy, noting that pineapple juice may help with sinus issues but works more slowly and less effectively than targeted treatments. Another expert might see it as a serious falsehood, arguing that promoting pineapple juice as a quick fix is misleading and potentially harmful. Both perspectives are valid, highlighting the inherent subjectivity in judging a claim's truthfulness. Experts' sensitivities can also vary depending on the topic, leading to apparently inconsistent judgments. These findings suggest that the existing practice—simply asking for a veracity label—needs to be redefined to align expert opinions in the first place.

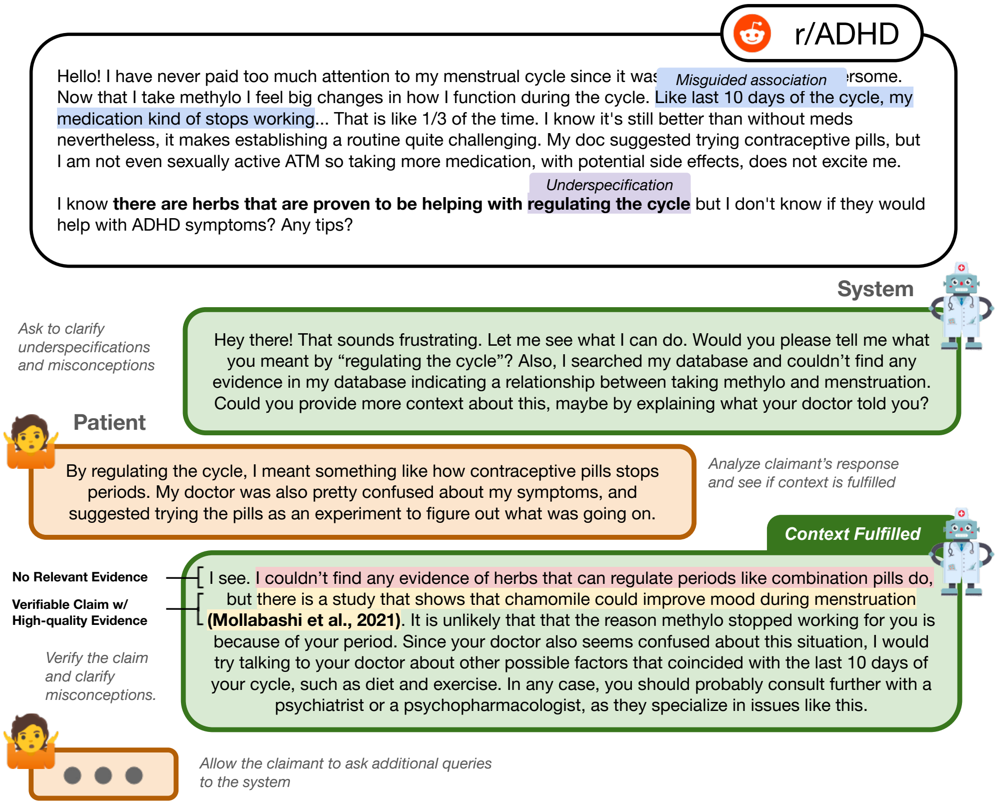
5 How Can We Address These Challenges?

To address the challenges in 4, we argue that fact-checking alone is insufficient: A **communication model** is required to mirror a dialogue between a healthcare provider and a patient. This frames fact-checking as a dialogue aimed at addressing patient's information needs. A vision of the user interaction with the communication model is shown in 2.

5.1 Clarifying Intent Through Conversation

In Section 4.2, we describe how naturally occurring claims often contain underspecifications that require understanding the patient's intentions. A communication model can address this through dialogue with the patient, similar to clinical interactions where physicians asking patients questions to understand their care needs. Prior work in dialogue systems has explored resolving ambiguity by generating clarification questions and modeling future conversations [Kim et al., 2023, Zhang et al., 2024a, Zhang and Choi, 2023].

To do this effectively, the system must identify underspecifications from the provided context. Similar work identifying ambiguities in user queries could provide a template for this task [Zhang et al., 2024b]. However, as we discovered in our analysis, identifying these underspecifications also requires extensive expert knowledge of the claim's subject matter, which could be encoded in a trained model or accessed via a retrieval-augmented system. The system must also identify "misguided" claims, as discussed in Section 4.2, which requires recognizing subtextual implications and commonly


r/ADHD
Misguided association
Underspecification
System
Ask to clarify underspecifications and misconceptions
Patient
Analyze claimant's response and see if context is fulfilled
Context Fulfilled
No Relevant Evidence
Verifiable Claim w/ High-quality Evidence
Verify the claim and clarify misconceptions.
Allow the claimant to ask additional queries to the system

Hello! I have never paid too much attention to my menstrual cycle since it was rsome. Now that I take methylo I feel big changes in how I function during the cycle. Like last 10 days of the cycle, my medication kind of stops working... That is like 1/3 of the time. I know it's still better than without meds nevertheless, it makes establishing a routine quite challenging. My doc suggested trying contraceptive pills, but I am not even sexually active ATM so taking more medication, with potential side effects, does not excite me.

I know **there are herbs that are proven to be helping with regulating the cycle** but I don't know if they would help with ADHD symptoms? Any tips?

Hey there! That sounds frustrating. Let me see what I can do. Would you please tell me what you meant by "regulating the cycle"? Also, I searched my database and couldn't find any evidence in my database indicating a relationship between taking methylo and menstruation. Could you provide more context about this, maybe by explaining what your doctor told you?

By regulating the cycle, I meant something like how contraceptive pills stops periods. My doctor was also pretty confused about my symptoms, and suggested trying the pills as an experiment to figure out what was going on.

I see. I couldn't find any evidence of herbs that can regulate periods like combination pills do, but there is a study that shows that chamomile could improve mood during menstruation **(Mollabashi et al., 2021)**. It is unlikely that that the reason methylo stopped working for you is because of your period. Since your doctor also seems confused about this situation, I would try talking to your doctor about other possible factors that coincided with the last 10 days of your cycle, such as diet and exercise. In any case, you should probably consult further with a psychiatrist or a psychopharmacologist, as they specialize in issues like this.

Figure 2: This figure illustrates the communication model for fact-checking, where the system engages the patient in dialogue—asking clarifying questions, filling contextual gaps, and verifying the claim and addressing any misconceptions.

held misconceptions. An ideal system would proactively query the patient to uncover incorrect assumptions and address them with empathy, fostering constructive engagement.

Direct communication with the patient is not strictly necessary to achieve intent clarification. Kim et al. [2023] proposed generating a tree of clarification questions, which, when fully answered, provides the context needed to resolve an ambiguous query. A similar approach could be applied here, where a tree of clarification questions resolves different interpretations of an underspecified claim, all of which must be verified and included in the system's final output. However, for this approach to work research is needed to study how to align the final output with the patient's original intent.

5.2 Guided Retrieval of Medical Evidence

In Section 4.1, we discussed the disconnect between what is practical and measurable in evidence-based medicine and what patients care about. The communication model enables providers to guide such claims toward verifiability while clarifying the patient's intent. To support this process, evidence retrieval should inform the dialogue between the patientand the system. When no relevant evidence is found, the system should communicate this and guide the patient toward related, verifiable claims. When none is available, the system should *abstain*. This approach makes clear that the claim is unverifiable while offering a pathway for continued learning.

This *guided retrieval* could also help correct misguided claims, as this conversational approach digs into and exposes the patient's thought process. Similar to physician-patient interactions, this process resembles a physician's response to an unanswerable query. They might first gather more information about the patient's question; if it remains unanswerable, they might recall related (answerable) questions. The field of Interactive Information Retrieval (IIR) studies the modeling and optimization of such back-and-forth interactions between users and retrieval systems [Zhai, 2020], e.g., for product retrieval [Wang et al., 2024, Aliannejadi et al., 2024]. Similarly, this approach could be used to guide unverifiable claims—often misguided due to false assumptions—toward claims that satisfy the patient's *information needs*.

5.3 Communicating Veracity Through Diverse Perspectives

As discussed in Section 4.3, our annotation study demonstrated that categorizing claims with fine-grained veracity labels is inherently subjective. While annotators often disagreed on labels, their reasoning in plain language explanations was often similar. During discussions, annotators often accepted each other's explanations as valid despite disagreeing on the level of severity. This suggests the possibility of a wider range of ostensible "agreement" can be reached. We propose that an effective medical fact-checking agent should produce responses that reflect diverse expert perspectives, acknowledging the inherent heterogeneity of expert evaluation. The need for these diverse explanations is corroborated by fact-checking professionals, who acknowledge the need and complexity of crafting thorough, nuanced explanations and calls for explanations to accommodate different audience needs [Warren et al., 2025]. Encouraging response diversity, rather than imposing artificial consensus via a single numerical value, is crucial for developing multi-agent medical fact-checking systems that integrate multiple expert viewpoints.

5.4 Medical Evidence Beyond RCTs

To bridge the gap between user questions and the limited pool of RCTs, future work could incorporate other forms of medical evidence such as meta analyses, cohort studies, case-control studies, case series, and case reports. Expanding the scope of medical evidence increases complexity, as systems must determine how to retrieve and synthesize evidence of various types and strengths, how to appropriately communicate this to end users. To this end, systems should adequately communicate its uncertainty according to existing guidelines in medicine [Ratcliff et al., 2021, Simpkin and Armstrong, 2019] and reference lower-grade evidence only when the communication model fails to identify a helpful, verifiable question and should clearly convey the quality of the source evidence and its limitations.

6 Conclusion

We have presented an analysis of the construct validity of existing end-to-end automatic medical fact-checking systems, with expert engagement in all key aspects of the system. Our work highlights the unique challenges of automated medical fact-checking, showing that it should be approached with user interaction in mind, and not as an end-to-end system. We proposed a communication model to clarify underspecifications and guide unverifiable claims, aiming to improve user outcomes and real-world utility. We hope this work inspires further exploration of human-in-the-loop systems for medical fact-checking.

Limitations Due to factors in time and cost, our annotation study is limited in the number of claims experts annotated. However, we were still able to collect a wealth of information from these annotations and from the insightful discussions with our expert annotators. Ultimately, the fundamental issues with the fact-checking process that we present in this paper acted as a barrier in continuing with a more large-scale annotation study. We used an automatic document retrieval system in lieu of a manual effort from annotators to find relevant abstracts in relation the claim. Therefore, it is possible for such a method to fail to identify the most relevant abstracts with respect to the claim. We did this to avoid overwhelming annotators and because the focus of the study is not on the retrieval of medical documents, but on the process of synthesizing such documents to verify a claim. In fact, a diversity of relevance in the abstract collection would have lent more insight into how such evidence is supposed to be synthesized. However, the fundamental issues we detail in this paper overshadowed and impeded in being able to do such an analysis.

Acknowledgments and Disclosure of Funding

We thank our medical experts Barry Wei, Gustavo Antonio Lima de Campos, Jon Stewart Hao Dy, Beatrix Krause-Sorio, Aislin Azcárate Kent, and Celina Bandowski for their annotation efforts.

This work is partially supported by Good Systems[2], a UT Austin Grand Challenge to develop responsible AI technologies, and National Institutes of Health (NIH) grant 1R01LM014600-01. We also acknowledge support from the National Science Foundation (NSF) grant 2211954 and the Wellcome Trust grant 313618/Z/24/Z.

References

Mohammad Aliannejadi, Jacek Gwizdka, and Hamed Zamani. Interactions with generative information retrieval systems, 2024. URL `https://arxiv.org/abs/2407.11605`.

Hilda Bastian, Paul Glasziou, and Iain Chalmers. Seventy-five trials and eleven systematic reviews a day: how will we ever keep up? *PLoS medicine*, 7(9):e1000326, 2010.

Sara Champlin, Michael Mackert, Elizabeth M Glowacki, and Erin E Donovan. Toward a better understanding of patient health literacy: A focus on the skills patients need to find health information. *Qualitative Health Research*, 27(8):1160–1176, 2017.

Jifan Chen, Aniruddh Sriram, Eunsol Choi, and Greg Durrett. Generating literal and implied subquestions to fact-check complex claims. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang, editors, *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 3495–3516, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.emnlp-main.229. URL `https://aclanthology.org/2022.emnlp-main.229/`.

Jifan Chen, Grace Kim, Aniruddh Sriram, Greg Durrett, and Eunsol Choi. Complex claim verification with evidence retrieved in the wild, 2024. URL `https://arxiv.org/abs/2305.11859`.

Li T Page MJ Chandler J Welch VA Higgins JPT Cumpston, M and J Thomas. Updated guidance for trusted systematic reviews: a new edition of the cochrane handbook for systematic reviews of interventions. *Cochrane Database of Systematic Reviews*, (10), 2019. ISSN 1465-1858. doi: 10.1002/14651858.ED000142. URL `https://doi.org//10.1002/14651858.ED000142`.

Anubrata Das, Houjiang Liu, Venelin Kovatchev, and Matthew Lease. The state of human-centered nlp technology for fact-checking. *Information Processing & Management*, 60(2): 103219, 2023. ISSN 0306-4573. doi: https://doi.org/10.1016/j.ipm.2022.103219. URL `https://www.sciencedirect.com/science/article/pii/S030645732200320X`.

Zhenyun Deng, Michael Schlichtkrull, and Andreas Vlachos. Document-level claim extraction and decontextualisation for fact-checking, 2024. URL `https://arxiv.org/abs/2406.03239`.

Susannah Fox and Maeve Duggan. Health online 2013. pew research center. *National survey by the Pew Research Center's Internet and American Life Project*, 2013.

Anisha Gunjal and Greg Durrett. Molecular facts: Desiderata for decontextualization in LLM fact verification. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen, editors, *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 3751–3768, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-emnlp.215. URL `https://aclanthology.org/2024.findings-emnlp.215/`.

Zhijiang Guo, Michael Schlichtkrull, and Andreas Vlachos. A survey on automated fact-checking. *Transactions of the Association for Computational Linguistics*, 10:178–206, 2022. doi: 10.1162/tacl_a_00454. URL `https://aclanthology.org/2022.tacl-1.11`.

Gordon Guyatt, Andrew D. Oxman, Elie A. Akl, Regina Kunz, Gunn Vist, Jan Brozek, Susan Norris, Yngve Falck-Ytter, Paul Glasziou, Hans deBeer, Roman Jaeschke, David Rind, Joerg Meerpohl, Philipp Dahm, and Holger J. Schünemann. Grade guidelines: 1. introduction—grade evidence profiles and summary of findings tables. *Journal of Clinical Epidemiology*, 64(4): 383–394, 2011. ISSN 0895-4356. doi: https://doi.org/10.1016/j.jclinepi.2010.04.026. URL `https://www.sciencedirect.com/science/article/pii/S0895435610003306`.

[2] `https://goodsystems.utexas.edu`

Gordon H Guyatt, Andrew D Oxman, Regina Kunz, Gunn E Vist, Yngve Falck-Ytter, and Holger J Schünemann. What is "quality of evidence" and why is it important to clinicians? *BMJ*, 336 (7651):995–998, 2008a. ISSN 0959-8138. doi: 10.1136/bmj.39490.551019.BE. URL https://www.bmj.com/content/336/7651/995.

Gordon H Guyatt, Andrew D Oxman, Gunn E Vist, Regina Kunz, Yngve Falck-Ytter, Pablo Alonso-Coello, and Holger J Schünemann. Grade: an emerging consensus on rating quality of evidence and strength of recommendations. *BMJ*, 336(7650):924–926, 2008b. ISSN 0959-8138. doi: 10.1136/bmj.39489.470347.AD. URL https://www.bmj.com/content/336/7650/924.

Naeemul Hassan, Bill Adair, James T. Hamilton, Chengkai Li, Mark Tremayne, Jun Yang, and Cong Yu. The quest to automate fact-checking. 2015. URL https://api.semanticscholar.org/CorpusID:79175.

Qisheng Hu, Quanyu Long, and Wenya Wang. Decomposition dilemmas: Does claim decomposition boost or burden fact-checking performance?, 2025. URL https://arxiv.org/abs/2411.02400.

Hasan Iqbal, Yuxia Wang, Minghan Wang, Georgi Nenkov Georgiev, Jiahui Geng, Iryna Gurevych, and Preslav Nakov. OpenFactCheck: A unified framework for factuality evaluation of LLMs. In Delia Irazu Hernandez Farias, Tom Hope, and Manling Li, editors, *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 219–229, Miami, Florida, USA, November 2024. Association for Computational Linguistics. URL https://aclanthology.org/2024.emnlp-demo.23.

Liqiang Jing, Ruosen Li, Yunmo Chen, and Xinya Du. FaithScore: Fine-grained evaluations of hallucinations in large vision-language models. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen, editors, *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 5042–5063, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-emnlp.290. URL https://aclanthology.org/2024.findings-emnlp.290/.

Prerna Juneja and Tanushree Mitra. Human and technological infrastructures of fact-checking. *Proc. ACM Hum.-Comput. Interact.*, 6(CSCW2), November 2022. doi: 10.1145/3555143. URL https://doi.org/10.1145/3555143.

Ryo Kamoi, Tanya Goyal, Juan Diego Rodriguez, and Greg Durrett. WiCE: Real-world entailment for claims in Wikipedia. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7561–7583, Singapore, December 2023a. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.470. URL https://aclanthology.org/2023.emnlp-main.470/.

Ryo Kamoi, Tanya Goyal, Juan Diego Rodriguez, and Greg Durrett. Wice: Real-world entailment for claims in wikipedia, 2023b. URL https://arxiv.org/abs/2303.01432.

Gangwoo Kim, Sungdong Kim, Byeongguk Jeon, Joonsuk Park, and Jaewoo Kang. Tree of clarifications: Answering ambiguous questions with retrieval-augmented large language models, 2023. URL https://arxiv.org/abs/2310.14696.

Jan-Christoph Klie, Richard Eckart de Castilho, and Iryna Gurevych. Analyzing Dataset Annotation Quality Management in the Wild. *Computational Linguistics*, pages 1–50, 07 2024. ISSN 0891-2017. doi: 10.1162/coli_a_00516. URL https://doi.org/10.1162/coli_a_00516.

Neema Kotonya and Francesca Toni. Explainable automated fact-checking for public health claims, 2020. URL https://arxiv.org/abs/2010.09926.

J Richard Landis and Gary G. Koch. The measurement of observer agreement for categorical data. *Biometrics*, 33 1:159–74, 1977. URL https://api.semanticscholar.org/CorpusID:11077516.

Iain J Marshall, Benjamin Nye, Joël Kuiper, Anna Noel-Storr, Rachel Marshall, Rory Maclean, Frank Soboczenski, Ani Nenkova, James Thomas, and Byron C Wallace. Trialstreamer: A living, automatically updated database of clinical trial reports. *Journal of the American Medical*

Informatics Association, 27(12):1903–1912, 09 2020. ISSN 1527-974X. doi: 10.1093/jamia/ocaa163. URL https://doi.org/10.1093/jamia/ocaa163.

Iain James Marshall, Veline L'Esperance, Rachel Marshall, James Thomas, Anna Noel-Storr, Frank Soboczenski, Benjamin Nye, Ani Nenkova, and Byron C Wallace. State of the evidence: a survey of global disparities in clinical trials. *BMJ Global Health*, 6(1):e004145, 2021. PMCID: PMC7786802.

Nicholas Micallef, Vivienne Armacost, Nasir Memon, and Sameer Patil. True or false: Studying the work practices of professional fact-checkers. *Proceedings of the ACM on Human-Computer Interaction*, 6(CSCW1):1–44, 2022.

Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen-tau Yih, Pang Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. FActScore: Fine-grained atomic evaluation of factual precision in long form text generation. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12076–12100, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.741. URL https://aclanthology.org/2023.emnlp-main.741/.

Jenny Moberg, Andrew D Oxman, Sarah Rosenbaum, Holger J Schünemann, Gordon Guyatt, Signe Flottorp, Claire Glenton, Simon Lewin, Angela Morelli, Gabriel Rada, Pablo Alonso-Coello, and GRADE Working Group. The GRADE evidence to decision (EtD) framework for health system and public health decisions. *Health Res. Policy Syst.*, 16(1):45, May 2018.

Isabelle Mohr, Amelie Wührl, and Roman Klinger. CoVERT: A corpus of fact-checked biomedical COVID-19 tweets. In Nicoletta Calzolari, Frédéric Béchet, Philippe Blache, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 244–257, Marseille, France, June 2022. European Language Resources Association. URL https://aclanthology.org/2022.lrec-1.26.

Liangming Pan, Xiaobao Wu, Xinyuan Lu, Anh Tuan Luu, William Yang Wang, Min-Yen Kan, and Preslav Nakov. Fact-checking complex claims with program-guided reasoning. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6981–7004, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.386. URL https://aclanthology.org/2023.acl-long.386/.

Dorian Quelle and Alexandre Bovet. The perils and promises of fact-checking with large language models. *Frontiers in Artificial Intelligence*, 7, 2024. ISSN 2624-8212. doi: 10.3389/frai.2024.1341697. URL https://www.frontiersin.org/journals/artificial-intelligence/articles/10.3389/frai.2024.1341697.

Chelsea L Ratcliff, Bob Wong, Jakob D Jensen, and Kimberly A Kaphingst. The impact of communicating uncertainty on public responses to precision medicine research. *Annals of Behavioral Medicine*, 55(11):1048–1061, 2021.

Arkadiy Saakyan, Tuhin Chakrabarty, and Smaranda Muresan. COVID-fact: Fact extraction and verification of real-world claims on COVID-19 pandemic. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli, editors, *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 2116–2129, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.165. URL https://aclanthology.org/2021.acl-long.165.

David L Sackett, William M C Rosenberg, J A Muir Gray, R Brian Haynes, and W Scott Richardson. Evidence based medicine: what it is and what it isn't. *BMJ*, 312(7023):71–72, 1996. ISSN 0959-8138. doi: 10.1136/bmj.312.7023.71. URL https://www.bmj.com/content/312/7023/71.

Mourad Sarrouti, Asma Ben Abacha, Yassine Mrabet, and Dina Demner-Fushman. Evidence-based fact-checking of health-related claims. In Marie-Francine Moens, Xuanjing Huang, Lucia

Specia, and Scott Wen-tau Yih, editors, *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3499–3512, Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.findings-emnlp.297. URL https://aclanthology.org/2021.findings-emnlp.297.

Julius Sim and Chris C Wright. The kappa statistic in reliability studies: Use, interpretation, and sample size requirements. *Physical Therapy*, 85(3):257–268, 03 2005. ISSN 0031-9023. doi: 10.1093/ptj/85.3.257. URL https://doi.org/10.1093/ptj/85.3.257.

Arabella L Simpkin and Katrina A Armstrong. Communicating uncertainty: a narrative review and framework for future research. *Journal of general internal medicine*, 34:2586–2591, 2019.

Achilleas Thoma and III Eaves, Felmont F. A brief history of evidence-based medicine (ebm) and the contributions of dr david sackett. *Aesthetic Surgery Journal*, 35(8):NP261–NP263, 07 2015. ISSN 1090-820X. doi: 10.1093/asj/sjv130. URL https://doi.org/10.1093/asj/sjv130.

Juraj Vladika, Phillip Schneider, and Florian Matthes. HealthFC: Verifying health claims with evidence-based medical fact-checking. In Nicoletta Calzolari, Min-Yen Kan, Veronique Hoste, Alessandro Lenci, Sakriani Sakti, and Nianwen Xue, editors, *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 8095–8107, Torino, Italia, May 2024. ELRA and ICCL. URL https://aclanthology.org/2024.lrec-main.709.

Ivan Vykopal, Matúš Pikuliak, Simon Ostermann, and Marián Šimko. Generative large language models in automated fact-checking: A survey, 2024. URL https://arxiv.org/abs/2407.02351.

Somin Wadhwa, Vivek Khetan, Silvio Amir, and Byron Wallace. RedHOT: A corpus of annotated medical questions, experiences, and claims on social media. In Andreas Vlachos and Isabelle Augenstein, editors, *Findings of the Association for Computational Linguistics: EACL 2023*, pages 809–827, Dubrovnik, Croatia, May 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-eacl.61. URL https://aclanthology.org/2023.findings-eacl.61/.

Mengzhao Wang, Haotian Wu, Xiangyu Ke, Yunjun Gao, Xiaoliang Xu, and Lu Chen. An interactive multi-modal query answering system with retrieval-augmented large language models, 2024. URL https://arxiv.org/abs/2407.04217.

Miriam Wanner, Seth Ebner, Zhengping Jiang, Mark Dredze, and Benjamin Van Durme. A closer look at claim decomposition. In Danushka Bollegala and Vered Shwartz, editors, *Proceedings of the 13th Joint Conference on Lexical and Computational Semantics (*SEM 2024)*, pages 153–175, Mexico City, Mexico, June 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.starsem-1.13. URL https://aclanthology.org/2024.starsem-1.13/.

Greta Warren, Irina Shklovski, and Isabelle Augenstein. Show me the work: Fact-checkers' requirements for explainable automated fact-checking. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, CHI '25, New York, NY, USA, 2025. Association for Computing Machinery. ISBN 9798400713941. doi: 10.1145/3706598.3713277. URL https://doi.org/10.1145/3706598.3713277.

ChengXiang Zhai. Interactive information retrieval: Models, algorithms, and evaluation. In Jimmy Huang, Yi Chang, Xueqi Cheng, Jaap Kamps, Vanessa Murdock, Ji-Rong Wen, and Yiqun Liu, editors, *Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval, SIGIR 2020, Virtual Event, China, July 25-30, 2020*, pages 2444–2447. ACM, 2020. doi: 10.1145/3397271.3401424. URL https://doi.org/10.1145/3397271.3401424.

Michael J. Q. Zhang and Eunsol Choi. Clarify when necessary: Resolving ambiguity through interaction with lms, 2023. URL https://arxiv.org/abs/2311.09469.

Michael J. Q. Zhang, W. Bradley Knox, and Eunsol Choi. Modeling future conversation turns to teach llms to ask clarifying questions, 2024a. URL https://arxiv.org/abs/2410.13788.

Tong Zhang, Peixin Qin, Yang Deng, Chen Huang, Wenqiang Lei, Junhong Liu, Dingnan Jin, Hongru Liang, and Tat-Seng Chua. Clamber: A benchmark of identifying and clarifying ambiguous information needs in large language models, 2024b. URL https://arxiv.org/abs/2405.12063.

A All Claims from Last Split

ADHD, Herbs, and Menstruation

Post (r/ADHD): Hello, menstruating people! How do your cycle and ADHD influence each other and how do you deal with it?
EDIT: After getting your responses I am reflecting again how medicine does not give a shit about women. It's truly insane. Thank you!
Hello! I have never paid too much attention to my menstrual cycle since it was never particularly bothersome. Now that I take methylo I feel big changes in how I function during the cycle. Like last 10 days of the cycle, my medication kind of stops working... That is like 1/3 of the time. I know it's still better than without meds nevertheless, it makes establishing a routine quite challenging. My doc suggested trying contraceptive pills, but I am not even sexually active ATM so taking more medication, with potential side effects, does not excite me.
I know there are herbs that are proven to be helping with regulating the cycle but I don't know if they would help with ADHD symptoms? Any tips?

Population: People with ADHD ,

Intervention: Herbs ,

Outcome: Regulating the menstrual cycle

Stimulants and Sodium

Post (r/ADHD): Stimulants vs. Sodium
Im wondering if anyone else has experienced this. I find that my stimulant medications (Adderall IR and Vyvanse) make me very sensitive to salt. If I have a higher sodium meal (eg ramen or canned soup, or even just mustard on my sandwich), I get very bloated. Its uncomfortable and lasts for a few days. Whenever I take a break from my meds, this doesnt happen. Ive had labs done for it in the past and it doesnt seem like anything medically problematic, but its uncomfortable and it really stresses me out.

Population: People with ADHD ,

Intervention: Stimulant medications (Adderall IR and Vyvanse) ,

Outcome: Sensitivity to salt (resulting in bloating)

Dandruff and Trikafta

Post (r/CysticFibrosis): Anyone had really bad flaky scalp or dandruff lately ? Think it could be due to trikafta. Could it be anything else

Population: Patients with Cystic Fibrosis (implied by the subreddit r/CysticFibrosis) ,

Intervention: Trikafta (a medication) ,

Outcome: Flaky scalp or dandruff

Pineapple Juice Reduces Inflammation

Post (r/CysticFibrosis): Anyone with sinus issues drinking pineapple juice?
It's a weird question, but I saw a post about pineapple juice being good for sinus issues (helps with the inflammation) and just wondered if anyone has done this? Some people were commenting about the high sugar content in pineapple juice not being good, but they get around that by taking a supplement instead of drinking the juice. Anyone?

Population: Patients With Cystic Fibrosis ,

Intervention: Pineapple Juice ,

Outcome: Reduced Inflammation/Fewer Sinus Issues

Trikafta and PMDD

Post (r/CysticFibrosis): Trikafta & PMDD
So, I believe trikafta has given me PMDD premenstrual dysphagia disorder. Every month, the week before my period I have extreme anxiety in a running dialouge in my head that is constantly negative. I've never been this way before. I also have horrible hormonal acne on my back & forehead which are very new to me as well.
My question is: any one else having this problem? My Dr said they are noticing a "negative interaction with estrogen and trikafta". Anyone find anything that helps??

Population: Patients with cystic fibrosis (implied by the Reddit thread r/CysticFibrosis), specifically females of reproductive age ,

Intervention: Trikafta ,

Outcome: Development of PMDD (premenstrual dysphoric disorder, not dysphagia disorder) symptoms, including extreme anxiety and hormonal acne.

Table 7: All claims from the last split. Claim in the post is highlighted. Given population annotation is highlighted in blue. Intervention is highlighted in pink. Outcome is highlighted in green.

We present all the claims from the last split of annotations we collected from medical experts in Table 7.

B Pilot Inter-evaluator Agreement

The initial agreement for the pilot set of 10 claims is shown Table 8.

Type	κ
Population	0.379
Intervention	0.355
Outcome	0.279
Overall	0.347
Tab Support	0.353
Overall Support	0.191

Table 8: Inter-evaluator agreement measured through Fleiss' κ for the training set of 10 claims.

C Refined Guidelines Agreement

The agreement for the refined guidelines is shown in Table 9.

Type	κ
Population	0.420
Intervention	0.131
Outcome	0.173
Overall	0.161
Tab Support	0.207
Overall Support	-0.011
Expert Support	-0.120

Table 9: Inter-evaluator agreement measured through Fleiss' κ for the updated guidelines test set of 5 claims.

D Why we Added PIO Elements?

After our pilot and refinement set, we realized that a source of disagreement in the annotators was because of the lack of PIO elements to focus on.

There was sometimes disagreement in cases where there are multiple PIO elements, leaving it to the discretion of the medical experts creates disagreement, as some choose to consider both the PIO elements or just focus on one. An example is shown in Table 10.

There are rare instances also where the annotated PIO highlights are incorrect, for example, accidentally labeling a population as an intervention. Therefore, we devise a LLM-based pipeline to extract the Population, Intervention, and Outcome elements from the post. Our prompt and expert validation are detailed in Appendix E.

Prednisone

Post (r/lupus): Cytoxan and prednisone
Rheumatologist says cellcept failed to protect my kidneys and now I have developed lupus nephritis.Im so upset. Prednisone messed up my hips so badly that they both need to be replacedI dont want to get back on it but rheumatologist says its to bring the inflammation down in my kidneys. Ive never been on Cytoxan but the side effects sound identical to a lupus flare. How am I supposed to be positive with news like this? I feel so defeatedI dont know what to do.

Expert 1: Overall consensus of relevant abstracts is that the combination of prednisone with cyclophosphamide (Cytoxan) is effective in treating kidney inflammation due to lupus nephritis (a3, a6, a8, a9, 10). This is in support of the original claim. However, one abstract found that kidney function continues to gradually deteriorate even with treatment (a2). Due to the majority of abstracts supporting the original claim however, the conclusion can be made that cyclophosphamide and prednisoine combination therapy is effective for decreasing renal inflammation in lupus nephritis.	**Expert 2:** The overall conclusion is that the claim is partially refuted. Prednisone alone appears to lead to renal deterioration (a2, a3, a6, a7, a8), but in combination with immunosuppressants, can have beneficial effects (a9). Abstracts a4 and a5 were somewhat relevant (did not specifically test prednisone effects). Irrelevant abstracts included a1 and a10. All relevant abstracts included lupus nephritis as the population. The person who made the claim appears to have arthritis plus lupus nephritis, therefore, none of the abstracts reported this exact population.	**Expert 3:** None of the given abstracts were relevant to verify the claim. None of the studies had a control group for prednison treatment in lupus nephritis. All groups in all of the given studies recieved prednison as a base treatment and compared this to the effects of an additional immunosuppressive drug. Since lupus (-nephritis) is an autoinflammatory disease, it is usually (depending on the severity) treated with immunosuppressive glucocorticoids such as prednison to inhibit the autodestruction of tissues and organs.
Expert 4: Overall, the abstracts partially support the use of prednisone to reduce kidney inflammation. With the exception of study a7, every other study included either prednisone or glucocorticoid in both the treatment and control groups. The difference usually is between glucocorticoid only or low-dose. Even a7, the only one that does not show a low dose of glucocorticoid use, might not appear to do so because methods may not be totally revealed in the abstract. Therefore, the abstracts suggest that this patient might receive a prescription for at least low-dose prednisone.	**Expert 5:** Core: The evidence supports the benefit of immunosuppressive medications such as Cyclophosphamide in addition to steroids and oral maintenance medications for those with lupus nephritis. Addendum: There is no study to support the superiority of Cyclophosphamide over other immunosuppressive medications especially if the patient has already had a poor response to other immunosuppressive medications such as Mycophenolate mofetil.	**Expert 6:** Abstracts a9, a3 a10, a6, a8 conclude that de combination of cytoxan and glucocorticoids shows better outcomes in patients with nephritis related to systemic lupus, which are two of the meds mentioned in the claim. The abstracts a7, a5, a2 are somewhat relevant. Since the majority of the abstracts did mentioned better outcomes using the medications from the claim, but cytoxan wasn't the only immunosuppressive drug compared in the studies, I'd say overall the abstracts partially support.

Table 10: This figure demonstrates how experts can misinterpret the claim to evaluate without explicit guidance on which PIO elements to focus on in the claim.

E PIO Extraction

We used automatic method of LLAMA 3.1 405B.

After preliminary prompt testing and consultation with our medical expert, we utilized PIO element definitions and questions in our prompt from Duke https://guides.mclibrary.duke.edu/ebm/pico:

PATIENT OR PROBLEM
How would you describe a group of patients similar to yours? What are the most important characteristics of the patient? Example: COVID patients, diabetics

INTERVENTION, EXPOSURE, PROGNOSTIC FACTOR
What main intervention, exposure, or prognostic factor are you considering? What do you want to do with this patient? Example: Remdesivir, Ozempic

OUTCOME
What are you trying to accomplish, measure, improve or affect? Example: pain, weight, 30 day mortality

Extract the Population, Intervention, and Outcome elements from the following claim from the following text. Write "None" if the element does not exist in the text.

Text posted by someone in Reddit thread r/[sub][sub_description]:

[post]

Highlighted claim:

[claim]

On a sample set of 55 samples (random 5 claims sampled per each of the 11 conditions), our medical expert from an esteemd hospital reviewed and reported % accuracy on the PIO extractions. He said it was ready to go, except for a few minor comments including the implied message as well as there were a few instances where the extracted Population and Intervention overlapped. He reported accuracy of above 90% for each element.

RCT Verifiability of the PIO elements doesn't necessarily mean RCT verifiability of the claim.

This was our validation for our PIO extraction and we used it to select claims and give PIO elements.

F Retrieval Strategies

For the first phase of this research of claims and abstracts we evaluated for the pilot, we used state-of-the-art embedding model, `Alibaba-NLP/gte-large-en-v1.5`, which was the top performing open-sourced embedding model with lightweight parameters on 06/03/2024 according to the HuggingFace MTEB Benchmark, which ensured feasible use to embed a database of 800,000 RCT abstracts with our limited resources.

For the next phase, we did more extensive experiments with `dunzhang/stella_en_400M_v5` evaluating different prompts and setups for retrieval.

We used SOTA embedding model, `dunzhang/stella_en_400M_v5`, which was the SOTA open-sourced embedding model with lightweight parameters on 09/11/2024 according to the HuggingFace MTEB Benchmark, feasible to use to embed a database of 800,000 RCT abstracts with our limited resources.

The overall scores are explained in Table 11

The claims are shown in Table 12

Strategy	**Query**	**Document**	**Pop.**	**Inter.**	**Out.**	**Overall**
S2P	Question PIO	PIO	10.9	8.1	9.8	7.4
S2P	PIO	Abstract	11.3	8.7	10.6	8.3
S2P	PIO	PIO	11.4	8.6	10.3	8.1
S2S	**PIO**	**Abstract**	**11.8**	**8.4**	**11.2**	**8.3**
S2S	PIO	PIO & Abstract	11.1	8.1	10.2	7.9
S2S	PIO	PIO	11.5	7.8	9.6	7.6
S2S	PIO	PIO & Title & Abstract	11.2	8.2	10.2	7.8
S2S	PIO	Title & Abstract	11.6	8.3	10.6	8.0

Table 11: Retrieval Strategies Results on Test Set of 5 Claims. 15 is full score: each claim is averaged across 10 abstracts. The score of 1 is irrelevant, 2 is somewhat relevant, and 3 is relevant.

Post	P	I	C	All
Grapefruit and siezures **Post (r/Epilepsy):** Grapefruit and siezures Guys, I am wondering if you have any issues or know about interactions between oxcarbazepine and/or levetiracetam and grapefruit? I believe it may make those medications work differently, but I am not sure. **Population:** Epilepsy patients (implied by the subreddit r/Epilepsy and the mention of seizure medications) , **Intervention:** Grapefruit consumption (in interaction with oxcarbazepine and/or levetiracetam) , **Outcome:** Medication efficacy (i.e., how the medications work)	20.7	11.1	19.3	10.4
COVID-19 **Post (r/MultipleSclerosis):** Anyone take Paxlovid? Hi everyone - one of my kids just tested positive for Covid so assuming I will as well in the next few days. I contacted my PCP and neuro who said that as soon as I test positive they would recommend getting me on Paxlovid as long as we can find it (vs monoclonal antibody infusions, as apparently those are not working quite as effectively against Omicron). Ive been doing my research on the drug but am curious if anyone has taken it yet. What was your experience? The worst side effect she mentioned was like a battery acid taste in your mouth for the entire 5-day regimen. Sounds gross but doable. Thanks in advance for any insight. **Population:** COVID patients (specifically those with Omicron variant, and implied to be similar to the author who has Multiple Sclerosis) , **Intervention:** Monoclonal antibody infusions (vs Paxlovid, but the claim specifically refers to the effectiveness of monoclonal antibody infusions) , **Outcome:** Effectiveness (implied to be against Omicron variant, but not a specific measurable outcome)	23.1	18.1	22.5	18.1
Balloon Sinuplasty **Post (r/Sinusitis):** After your balloon sinuplasty, how long did it take you to get rid of your face pressure and shallow breathing? Its been 5 days since my procedure and I still have facial pressure. This effects me in negative ways like not being able to focus with my eyes. I also have trouble automatically deep breathing. Im just curious how long does it typically take to recover from these symptoms? I was told I should experiencing some relief after 2 days but still nothing. Im starting to wonder if this is dental related. **Population:** Patients who have undergone balloon sinuplasty (similar to the author of the post, who has sinusitis) , **Intervention:** Balloon sinuplasty , **Outcome:** Relief from facial pressure and breathing symptoms (specifically within 2 days after the procedure)	8.3	8.2	8.7	7.9
Elvanse **Post (r/ADHD):** Picking the correct meds? So I'm UK based and have just started titration through PsychiatryUK after waiting nearly a year. Started on Xaggitin XL 18mg (now on 36mg and then will go to 54mg). It seems to work ... I wake up a lot easier as before I would wake up groggy and just completely unfunctional .. which would then lead to me being easily distracted etc, and my focus throughout the day seems improved ... mood seems improved ... but, at around 4pm I crash hard and just feel exhausted at night. I had some family over at the weekend and was unable to keep up in conversation. Also after a few days it seems to have less of an effect ... but then I up the dose and it works again. Obviously this won't be sustainable and I will eventually reach max dose. I would like to try Elvanse next as I've heard people say the come down is less brutal ... but it takes longer to kick in, which is a problem for me as I find it most difficult in the morning. I could wake earlier and take it, and then go back to bed ... or I've also seen some people who take Amfexa as a top up in the morning and then take Elvanse for the rest of the day. This seems totally ideal for me, and has the benefit where I can forgo the meds on weekends if I'm not working / doing much - but have the quick release to fall back on in the event something crops up. The question is - can I request this? I scared to accept a med just because it works somewhat, as changing it seems like it would be a pain in the ass .. so I'd like to make an informed decision.. but also don't want to waste their time. I'm also scared to even mention Amfexa for fear of being labeled as someone just wanting a high. **Population:** ADHD patients (specifically, those similar to the author who experience a crash with their current medication) , **Intervention:** Elvanse , **Outcome:** Severity of the "come down" (i.e., the crash or comedown experienced by the patient after the medication wears off)	18.4	15.7	14.6	13.1

Continued on next page

Post	P	I	C	All
Probiotics IBS	20.3	13.1	17.4	13.9
Post (r/ibs): Some Info on Probiotics for IBS So I posted this in a comment and it was suggested that I make this info its own post. I did a bunch of research into what probiotics have shown to be effective for IBS in various research studies. From the links below, the collective summary is that these probiotic strains (alone or in combination with others) were effective in one or more studies in reducing one or more IBS symptoms in frequency and/or severity: Saccharomyces Boulardii CNCM I-745 (actually a yeast), Bifidobacterium longum 35624, Bifidobacterium infantis, Lactobacillus acidophilus, Lactobacillus plantarum 299v, Lactobacillus rhamnosus GG, Saccharomyces cerevisiae CNCM I-3856, Bifidobacterium lactis HN019, Bifidobacterium lactis BB-12, Lactobacillus acidophilus NCFM, Bifidobacterium lactis Bi-07, and Bifidobacterium bifidum MIMBb75. I may have missed a few in the above list, but the main gist is that Bifidobacterium and Lactobacillus species probiotics, as well as Saccharomyces Boulardii, are the main researched probiotics that show some kind of improvement for IBS (though it also depends on the type of IBS, and most of them helped with one or some symptoms but not all of them - see links below for more detail). Also, many of the studies couldn't draw conclusive evidence because there just wasn't enough data/research on specific probiotics yet, or because the studies didn't fulfill certain criteria. But they were able to state what evidence they did have, and identify what is likely to be the case based on findings so far. Bifidobacterium longum 35624 used to be called Bifidobacterium infantis 35624 and is still sometimes referred to that way. It is patented by Align, which is the brand name for that probiotic. Floraster is a leading brand name for Saccharomyces Boulardii CNCM I-745. There are other good probiotic brands too though (that have similar or different probiotic strains). A lot of this will be trial and error as you find what works for you. Be sure to give each probiotic or probiotic blend you try at least a 4 week trial, since it takes a bit of time for your body to adjust and for the probiotics to get to a certain level in your system to start working. It's a good idea to make sure that the probiotics are third-party tested and made in a GMC (or equivalent) registerred facility, to be sure you are getting exactly what's on the product label. I'm not sure exactly how it works in other countries, and that advice is US-based, but some level of quality assurance is generally advised. Also, for those who aren't used to reading research papers like this, usually the important information is in the Results and/or Discussion section. Sometimes they put a summary of those near the top, and sometimes you have to scroll to the end of the article. It's also a good idea to read the abstract/summary at the top of the article so you have an idea of what the paper is about. Reading the other sections of the paper will give you more detail into what they did for the study, which will give more context, but isn't usually necessary to read the whole thing to understand the end results, which tell you what they found about the probiotics in relation to IBS. And of course, always talk to your doctor about medical stuff! I'm just providing this information in case it might be helpful. I am NOT giving medical advice or telling you all to take probiotics or endorsing any brands or types or anything. Whether you decide to take probiotics or not is up to you and your doctor. This is just an info dump in case anyone wanted to know this stuff. Here are the links: https://www.thelancet.com/journals/eclinm/article/PIIS2589-5370(21)00434-X/fulltext https://www.sciencedirect.com/science/article/pii/S1590865819309594 https://www.ncbi.nlm.nih.gov/labs/pmc/articles/PMC7279071/ https://www.optibacprobiotics.com/learning-lab/in-depth/gut-health/which-probiotics-are-best-for-ibs#IBS-U https://www.frontiersin.org/articles/10.3389/fphar.2020.00332/full#B68 https://badgut.org/information-centre/a-z-digestive-topics/probiotics-for-irritable-bowel-syndrome/ https://www.ncbi.nlm.nih.gov/labs/pmc/articles/PMC3296087/ https://pubmed.ncbi.nlm.nih.gov/33745570/ https://onlinelibrary.wiley.com/doi/abs/10.1111/eci.13201 **Population:** Patients with IBS (Irritable Bowel Syndrome) , **Intervention:** Probiotics (various strains) , **Outcome:** Reduction in IBS symptoms (frequency and/or severity)				

Table 12: Retrieval Test Ratings (Population, Intervention, Outcome, Overall) by Claim. Full Score is 24. Claim in the post is highlighted. Given population annotation is highlighted in blue. Intervention is highlighted in pink. Outcome is highlighted in green.

G Annotation Guidelines

We present the guideline given to our expert annotators below. These instructions were given in slide-deck format to annotators with images from the annotation interface spliced in-between to clearly indicate how to annotate.

The Task
This annotation task involves verifying medical claims made on Reddit posts using retrieved evidence. You will be looking at the provided abstracts to determine whether, when considering all the evidence, you can support or refute the claim.

Post
We will give you a Reddit post, which is annotated to contain the following.

- What subreddit the post is from.
- Spans indicating PIO (Population, Intervention, Outcome) elements.
  - Population indicates the affected subjects (ex: COVID patients, diabetics).
  - Intervention indicates any treatments applied to the subjects (ex: remdesivir, Ozempic).
  - Outcome indicates how the effects of the intervention are evaluated (ex: pain, weight, 30 day mortality).
- Claim Span: Part of the post that makes the medical claim that we analyze.

Post & Derived Claim

- You will NOT be directly evaluating the information in the post. It is presented to you as to inform you of the context in which the claim is made.
- What you will be directly evaluating is the claim derived from the post. We present you with a (P, I, O) tuple extracted from the post that we use to make the claim as clear as possible.
- In some cases, the claim in the post may be ambiguous. In this case, we will present a disambiguated claim for you to evaluate.

RCT-Verifiability

- A claim is RCT-Verifiable if there exists (or should exist) a reasonable RCT that will be able other either support or refute it.
  - A reasonable RCT is one that can be practically and ethically conducted.
- Most of the claims we give should be RCT-Verifiable. However, this may not always be the case.
- In the case when a claim is not RCT-Verifiable, you should indicate as such.
  - You will be forced to write a 10 word explanation for why the claim is not RCT-Verifiable. Please ensure that the explanation is for a legitimate reason, for example, an unethical intervention, as we will review. You can only continue to annotate once you are done.

Retrieved Abstracts

- For each claim, you are given 10 abstracts that are retrieved automatically based on information in the claim.
- Each abstract has the following:
  - Title
  - Published Date
  - Informative Highlights: PIO Spans and Abstract Punchline (Span describing the core of the abstract's findings)
- We provide you a way to flag abstracts that you believe to be of poor quality in the interface. Be sure to keep in mind the quality of the RCT experiment described in the abstract when annotating them.

Relevance Annotations

- For these annotations, you will be analyzing the relevance of the abstract to the claim being evaluated.
- You will analyzing the relevance of the following four components:
  - Population: Is the population being studied in the abstract relevant to the population the claim is addressing?
  - Intervention: Is the intervention being studied in the abstract relevant to the population the claim is addressing?
  - Outcome: Are any of the outcome measures used in the RCT described in the abstract relevant to the population the claim is addressing?
  - Overall: Is the abstract relevant enough to the claim for it to be used to verify the claim?
- As mentioned before, you may flag the abstract if you think it is of concerning quality.

Relevance Labels

- For each relevance component, you are given 4 labels to choose from. They are as follows:
  - Select (Default)
    - For PIO: The element is missing.
    - For Overall: The abstract does not describe an RCT.
  - Irrelevant
    - For PIO: The element in the abstract has no relation at all to the corresponding element in the post.
      - Ex for Population - Claim: Patients with LPR - Abstract: Healthy Patients
    - For Overall: No part of this abstract can be used to make even an inference on whether the claim can be supported or refuted.
- Somewhat Relevant
  - For PIO: Indicates that the element has some relation to the corresponding element in the post, but is not close enough for it to be used to directly verify the claim even if all other elements are 100% relevant.
    - Ex for Population - Claim: Patients with LPR - Abstract: Patients with gastro-oesophageal reflux disease
  - For Overall: Some parts of this abstract can be used to make an inference on whether the claim can be supported or refuted. However, this abstract still cannot be used as direct evidence to support or refute the claim.
- Relevant
  - For PIO: Indicates that the element in the abstract is close enough to the corresponding element in the post for the purpose of verifying the claim.
    - Ex for Population - Claim: Patients with LPR - Abstract: Patients with laryngopharyngeal reflux
  - For Overall: The abstract can be used as direct evidence to support or refute the claim.

Abstract Support

- If overall, the abstract is relevant to the claim. You will be given the opportunity to annotate for whether the abstract supports/refutes the claim.
- There are four labels you can choose from:
  - Refutes: This abstract fully refutes the claim in the post.

    - Partially Refutes: This abstract refutes the claim given some condition or caveat.
    - Partially Supports: This abstract supports the claim given some condition or caveat.
    - Supports: This abstract supports the claim in the post.
- A partial support or refute indicates that there is some nuance in the RCT result that prevents the abstract from fully supporting or refuting the claim.
    - A sub-group of the population experienced different results from the rest.
    - The results can only be reproduced under specific conditions that cannot be generalized.

Relevant Span
When you are done determining the support label for abstract. You must determine which span of text in the abstract is most relevant in indicating whether the abstract can be supported or refuted.

Tiering

- After you are done with annotating all the abstracts you can start the tiering and synthesis.
- In the tiering phase, you are organizing the abstracts into tiers.
    - Abstracts are automatically tiered according to the relevance annotations.
    - You should attempt to further categorize the abstract according to their quality or their importance regarding the claim, as well as temporal relevance (up your own medical expertise).

Synthesis

- For this task, you must pick the label determining whether the claim is supported or refuted according to two criteria.
    - Overall Support (OS): Determine whether the claim is supported or refuted using only the provided evidence.
    - Expert Opinion (EO): Determine whether the claim is supported or refuted using your expert knowledge.
- You are given 6 labels to choose from:
    - No Relevant Abstracts/No Expert Opinion:
        *
        * OS: There are no relevant abstracts to determine overall support.
        * EO: You don’t have the expert knowledge in the field to make this decision.
    - Refutes:
        * OS: Overall, considering all the abstracts, there is strong evidence that the claim can be refuted.
        * EO: According to your expert knowledge, this claim can be strongly refuted

      .
    - Partially Refutes:
        * OS: Overall, considering all the abstracts, there is evidence that the claim can be refuted depending on some general condition or caveat.
        * EO: According to your expert knowledge, this claim can be refuted depending on some general condition or caveat.
    - Inconclusive:
        * OS: This should rarely happen. Only pick this in cases, where there is true deadlock within the evidence as to whether the claim can be supported or refuted.
        * EO: According to your expert knowledge, there is no scientific consensus that points to the claim being supported or refuted.
    - Partially Supports:
        * OS: Overall, considering all the abstracts, there is evidence that the claim can be supported depending on some general condition or caveat.
        * EO: According to your expert knowledge, this claim can be supported depending on some general condition or caveat.
    - Supports:
        * OS: Overall, considering all the abstracts, there is strong evidence that the claim can be supported.
        * EO: According to your expert knowledge, this claim can be strongly supported.

Synthesis Explanation

- Afterwards, write an explanation of why you picked that option. Use the tiers you created earlier to help develop this explanation (quality, temporal, relevance). Cite the abstracts in your explanation.
    - For example: There were a few abstracts that refuted the claim (a2, a5). ...
- Give sufficient context and details where someone can follow the reasoning without looking at your annotations. Think of the perspective of you explaining to a patient.
- Try and keep the grade level at around Middle School, try to avoid complex jargon
- We would like for you to include the following in your explanation:
    - Main statement explaining why you selected the label.
    - Rundown of how relevant evidence (abstracts) supports/conditionally supports/refutes claim.
    - (optional) Addendum with relevant clinical experience regarding claim
    - Use the terminology of a(abstract number) like a9 to refer to abstract 9 in your explanations.
    - Aim for the length (without addendum) to be around 100 words. If you don’t need that much explanation, 50 words is fine. If you really need to explain something in more detail, please keep it under 150 words.
- Please do not include:
    - Any direct references to the tiers (Ex: The abstracts in tiers 1).

H Annotator Recruitment

We recruited five of our experts from Upwork, a process that took four weeks. During this phase, we received 117 proposals on Upwork, and reached out to 19 people with a sample annotation task for claim in 13 to gauge their instruction-following capability and and their medical expertise through the quality of their annotation and plain language explanations. From this, we selected 7 of the most qualified candidates by explanation quality, interviewed them for final fit, and ultimately chose 5 which best met our expectations. Our experts worked anywhere between 3 to 20 hours per week of annotation. Our medical experts on average took 20 minutes to annotate each claim end-to-end. We paid our experts a range between $22 and $35 through Upwork.

Our total spend for the annotation study was $1432.20.

Gaviscon Advance

Post (r/GERD): Can I buy liquid alginate suspension (Gaviscon Advance) in the U.S.? Hi everyone. I'm newly diagnosed with LPR and doing a lot of research on the best treatments. I've read that a liquid alginate suspension (Gaviscon Advance) is quite effective at treating LPR but it looks like it's not sold in the U.S. Does anyone know how I can find it here?

Table 13: The claim is highlighted. The RedHOT intervention annotation is highlighted in pink.

I Pilot Claims

Post

microclots

Post (r/Diabetes): Could microclots help explain the mystery of long Covid? Acute Covid-19 is not only a lung disease, but actually significantly affects the vascular (blood flow) and coagulation (blood clotting) systems. A connection to the damage done by diabetes might be possible.

Diabetes

Post (r/Diabetes): Affording Medication
so im on a family plan with a 3k/6k out of pocket expense, I think it's hdhp with a hsa that my husband employer contributes a bit too. I know when I had coverage with my job i had a ppo plan. he's the one that chooses the plans at his job so im not the best when it comes to explaining the details for it..
was orginally taking metformin but it's horrible and over the past 2 months it's been making me sick as a dog so I asked my endocrinologist can I go on something else. she recommended ozempic since alot of patients responded well to it, lost weight, and had a good effect on their sugar. plus it's taken only weekly in which sounds great for someone like me since I'm not the best with keeping up with medications. back in December since we had hit our 6k deductible I had paid nothing when I recieced the medication so I had no clue what the actual price would be but I nearly had a heart attack when I tried picking it up in the store recently...with my plan I'm at 800 bucks for the thing and optum informed me it's 2300 (1981 with the discount card) for a 90 day supply. that's ALOT of money...I was going to purchase farxiga today with optum (1500 dollars) but literally don't have the money to afford to do so..my car needs a new catalytic converter so finacially I had to make the cut to my medication (my cardiologist put me on that to prevent heart failure since I have "resistant hypertension" that's not responding well medication)
i made a joke to my husband and said I may have to divorce him just so I qualify for that government health insurance. hell looking at it now I may be serious! as a diabetic or anyone in America on any type of medication how are ppl able to afford their insulin/pills/machines/ whatever. our household income is around 85k so there's not much assistance we can get that im aware of

Hallucinations

Post (r/narcolepsy): Do people with IH experience hallucinations?
I am so confused! My MSLT showed IH but my doctor gave me a clinical diagnosis of narcolepsy because I experience hypnopompic hallucinations and sleep paralysis. She told me people with IH dont experience those things which is why she switched the diagnosis. Im confused because Ive read articles that say they are symptoms of IH. I know it doesnt really matter because treatment is the same, but I have this thing in me where I just need to know.

COVID

Post (r/Epilepsy): Epilepsy Patients Much More Likely to Die of COVID

Long Covid

Post (r/CFS): I Had Never Felt Worse: Long Covid Sufferers Are Struggling With Exercise And experts have some theories as to why. - The New York Times

Glycemic

Post (r/Diabetes): Dietary carbohydrate restriction augments weight loss-induced improvements in glycaemic control and liver fat in individuals with type 2 diabetes: a randomised controlled trial. (Pub Date: 2022-01-07)

Pfizer vaccine

Post (r/CysticFibrosis): Pfizer vaccine
My son, non cf, is having his second pfizer vaccine. He is 25 yrs old. For some reason I'm really nervous about it as he has been told not to exercise for 48hrs afterwards due to heart inflammable young people are getting...obvs this is rare...but my son is extremely active & I'm in a tizz. He's having now as i write this. I'm extremely proud he is having it as alot of youngsters are refusing it atm but the anxiety over it is making me feel sick.

mold

Post (r/rheumatoidarthritis): Mold and RA
I'm having a bit of a weird issue with mold. I'm currently in the process of being diagnosed with RA. I've got achy joints, swelling whole nine yards. I transferred job locations earlier this month and was starting to feel better and my hand swelling finally went down. I then signed up for some overtime in my old job location and after about 2 hours my elbows and hands started to ache and swell. Every time it rains at this building water runs through the walls. I'm certain theirs mold in the walls. Google says long-term toxic mold exposure can mimic RA. Had anyone else had an experience with RA symptoms not ending up being RA or having one large trigger to RA symptoms. After going home and sleeping on things my hands started to feel better but not completely.

Continued on next page

Post
Copaxone **Post (r/MultipleSclerosis):** copaxone vs aubagio? My gf is about to switch from once a day copaxon injections to aubagio at the advice of her new neurologist. After doing some research before starting the treatment, she is a bit worried about the liver function concerns with the drug. My gf is bipolar, has high anxiety, and is on several meds for her mental health. I just pulled up a site that compared these 2 drugs and was really angry to see that copaxone patients reported it caused depression, anxiety, and other things the doctor never mentioned. So I am cautiously optimistic that the change is in her best interest. Any thoughts or experiences would be greatly appreciated. My research seems to lean towards the new medication, but we are obviously concerned at least about the liver function monitoring. tyvm
Calcium **Post (r/thyroidcancer):** Calcium supplements (Citracal slow release) and total thyroidectomy Hey all, I'm almost 6 years post total thyroidectomy, and since my providers at the time of my TT didn't really share any of this info/it was hard to track down, I wanted to put it out there for others. First– you'll probably want to get on a calcium supplement. That part I was told. How much calcium I wasn't told, but eventually found out from a pharmacist to go a bit above the recommended for your age/assigned sex at birth. Normally my recommended would be 1000mg, but because of the TT, it's 1200mg. Second– wait at least 4 hours after taking your thyroid hormone replacement before taking a calcium supplement. Also was told that, also something everyone here probably already knows. Third– our bodies can only absorb around 500mg of calcium in one go. Fourth– if you've had a TT, your body will absorb calcium citrate more effectively than calcium carbonate. I learned this literally a month ago from a PCP who doesn't specialize in thyroid health, and I'd love to know why my endocrinologist never told me. Now, all of the above led me to be interested in the Citracal slow release, as it's 1200mg, but released slowly so you can take it once a day and still get all of it. My only issue was that I couldn't find anywhere that said how long it took to fully release. I was worried that if it took too long, it would prevent my levothyroxine from absorbing the next day. I couldn't find the answer online, but finally called their questions line today and found out it's 8 hours. Obviously I'm not a medical provider, I just want more of us to have access to this info, especially since a lot of us have worked with medical providers that don't give us all of the information we need. Also not here to advertise for that brand specifically, I just wanted something convenient enough to take once a day, and figured others might have had the same question!

Table 14: Ten Pilot Claims. Claim in the post is highlighted. Given population annotation is highlighted in blue. Intervention is highlighted in pink. Outcome is highlighted in green.

J Refinement Claims

Post
Ivabradine **Post (r/POTS):** Anyone here with low bp take Ivabradine? Im just wanting to do a bit of research on different meds before my doctors appointment. Last time they told me they cant medicate me because my bp dropped pretty low during the TTT. However, at rest my bp is normal and even standing up it doesnt drop noticeably low unless Im standing still for a longer period of time. So I just wanted to know if any of yall are in a similar situation and have good (or bad) experiences with this drug. I hear midodrine is good for low bp but its more expensive and the side effects sound kind of iffy to me.
Fluoxetine **Post (r/Dysthymia):** Long-lasting apathetic tendencies, anhedonia etc. I'm just apathetic in general, and am unable to do even the smallest things. Fluoxetine might have had some positive effects, and I'm supposed to be taking it now, but I can't even be bothered to get a refill. I can't tell whether or not my asocial tendencies are a personality trait. I currently have no interest in maintaining a relationship with family or friends. I've never been diagnosed with dysthymia - only depression - but a lot of the symptoms seem relevant, and my doctor did mention it at one point.
Psychosis and Antidepressants **Post (r/Psychosis):** Psychosis and antidepressants Hey everyone! So some crazy stuff happened to me over the last week. I am on abilify for my psychosis and I have been suffering from depression. My doctor decided to prescribe me Wellbutrin 150mg first. Took it for about five days, started having extreme anxiety and dry mouth. I mentioned this to my doctor and he switched me to Lexapro 5mg. Extreme anxiety and dry mouth but something new happened this time -my fucking delusions and hallucinations came back. I had to legit tell myself my thoughts werent based on reality. But holy crap was it difficult. I didnt take any antidepressants today and already feel better. This is crazy, has anyone experienced anything like this? I didnt think anti depressants would bring out my psychosis. Guess I might have to go the natural route for my depression :(

Continued on next page

Post
Prednisone **Post (r/lupus):** Cytoxan and prednisone Rheumatologist says cellcept failed to protect my kidneys and now I have developed lupus nephritis.Im so upset. Prednisone messed up my hips so badly that they both need to be replacedI dont want to get back on it but rheumatologist says its to bring the inflammation down in my kidneys. Ive never been on Cytoxan but the side effects sound identical to a lupus flare. How am I supposed to be positive with news like this? I feel so defeatedI dont know what to do.
Metformin replace Insulin **Post (r/Diabetes):** Can metformin replace insulin? I realize this is definitely case-by-case but Im curious to know if anyone has been able to get off of insulin and take just metformin? When I was diagnosed with type 2 I was automatically put on insulin and generally take quite a bit of it, but now after some research Im considering asking my doctor to try treatment through metformin in February.

Table 15: Retrieval Test Ratings (Population, Intervention, Outcome, Overall) by Claim. Full Score is 24. Claim in the post is highlighted. Given population annotation is highlighted in blue. Intervention is highlighted in pink. Outcome is highlighted in green.

K Automatic Pipeline for Claim RCT Verifiability

We first started with this prompt for RCT Verifiability.

Can the following claim from the text from r/subreddit be verified by conducting a randomized controlled trial? To be verifiable the claim must have a clear intervention that would be ethical to perform given the context. Give a classification of verifiable or not verifiable.

Claim: [claim]

Text from r/[subreddit]: [post]

On a small set, our medical expert did not approve.

Our medical expert proposed this refined prompt with a persona-style instruction to be a potential clinical trialist as well as more detailed ethical guidelines:

You are a potential clinical trialist. I will give you a claim and post. The claim is part of the post, and the post can give you context. I want you to tell me if the claim can be studied in a randomized controlled trial (RCT). An RCT can test an intervention to measure a benefit or non-inferiority. However, the RCT must be ethical: Ethical Guidelines: The intervention should not cause harm or have a significant risk of toxicity. It should not test exposures known to be potentially harmful, such as food-drug interactions that might cause adverse effects. The safety of participants is the primary concern, and interventions that pose significant health risks should not be tested in an RCT. Design Requirements: The trial needs to have a control group, with the only difference being the intervention. There must be a feasible and ethical way to measure outcomes without exposing participants to undue risk. Wait for my text to classify whether the claim can be ethically studied in an RCT.

Claim: [claim]

Text from r/[subreddit]: [post]

Format your response starting with Classification: Can be ethically studied in a RCT or Classification: Cannot be ethically studied in a RCT

L Implementation

We emphasize that the communication model is an abstraction, and that the actual system could be implemented in many ways. The model could be a Reddit or Chrome extension tool for users to interact with while they write their Reddit post. Alternatively, the model could also publicly post follow-up questions as comments as a public reply to the users' post. The benefit of this approach is that there could be an expert in the loop that verifies the models' response, as well as these public responses could be helpful to other users reading the thread. We encourage the research community to explore various implementations of this system, as well as focus on extensive human and expert evaluation and systematic HITL methods.

M Plain Language Explanation Guideline

- Include an overall sentence either at the beginning or end of your synthesis explanation.
- Target to aim the explanations at 100 words or less, 150 words if there are details that must be elaborated on.
- Include details of abstracts identified as relevant and explanations of how it supports the ultimate label, including some nuance.
- (Optional) Medical Addendum at end.

N Changes between annotation rounds

In the first round, ten claims were annotated without PIO claim contextualization and the expert support field. In the second round, we adjusted the annotation guidelines to clarify labels to increase agreement. The resulting agreement was still low. After receiving feedback from annotators, we revamped our system by adding a filter to filter non-RCT verifiable claims, extract and provide PIO elements for the annotators to focus on, improved the retrieval system, and further clarified the annotation guidelines. For the third round, experts annotated five claims with these new updates.

O Ethical Concerns

The posts found within the RedHOT dataset do contain health-related comments that are inherently sensitive. To respect this sensitivity, the authors of the RedHOT dataset notified all users of their inclusion in this dataset and provided them with the opportunity to opt-out. They also did not release the data directly, but instead provided a script to download content from Reddit so that individuals may be able to remove their post in the future. In this work, we do directly release a small subset of these posts from RedHOT that we used in our annotation study. In our released data, we do not reveal the username of the author of the post. We only include the text from the post and information about the subreddit in which it was found. Considering the measures taken by the authors of the RedHOT dataset and the fact that these posts have been publicly available on Reddit for at least more than an year, we believe it is safe to publicly release this data.

We have consulted with our Institutional Review Board (IRB) about the nature of our work and confirmed that the use of the RedHOT data and the subsequent annotation study using these data do not constitute research of human subjects. However, we do acknowledge that certain uses of this data may be considered sensitive. We strongly encourage researchers to obtain prior approval from their own IRB regarding the intended use of the data released by this work.

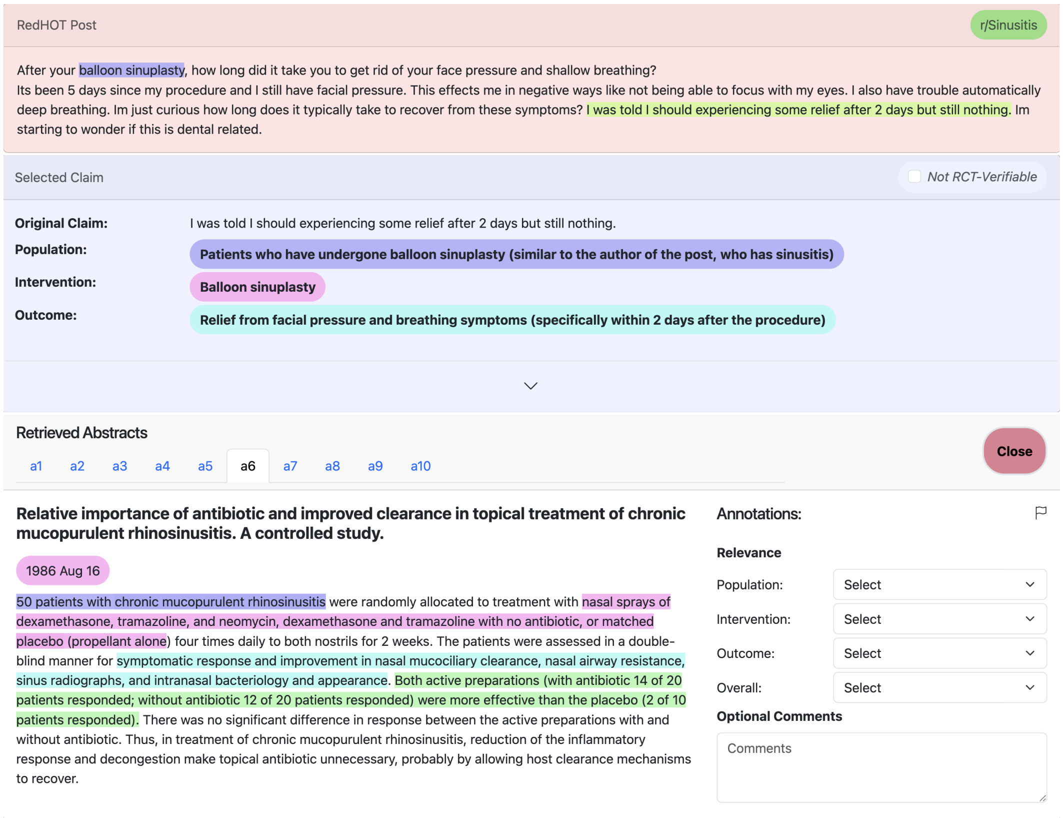
P License Information

We predominantly used the RedHOT dataset and abstracts from the TrialStreamer database in our work. Both of these works are licensed on a Creative Commons Attribution 4.0 International License.

We will also release our work under a Creative Commons Attribution 4.0 International License.

Q Annotation Interface

We used a web-based annotation interface to collect annotations from our expert medical annotators. Figure 3 shows how we present a claim with its surrounding context, extracted PIO elements for that claim, and retrieved abstracts corresponding to that claim. In this setup, we highlight the claim within the post in which its found along with any PIO spans as determined by data found in the RedHOT



RedHOT Post

r/Sinusitis

After your balloon sinuplasty, how long did it take you to get rid of your face pressure and shallow breathing?
Its been 5 days since my procedure and I still have facial pressure. This effects me in negative ways like not being able to focus with my eyes. I also have trouble automatically deep breathing. Im just curious how long does it typically take to recover from these symptoms? I was told I should experiencing some relief after 2 days but still nothing. Im starting to wonder if this is dental related.

Selected Claim

□ *Not RCT-Verifiable*

Original Claim: I was told I should experiencing some relief after 2 days but still nothing.

Population: **Patients who have undergone balloon sinuplasty (similar to the author of the post, who has sinusitis)**

Intervention: **Balloon sinuplasty**

Outcome: **Relief from facial pressure and breathing symptoms (specifically within 2 days after the procedure)**

Retrieved Abstracts

Relative importance of antibiotic and improved clearance in topical treatment of chronic mucopurulent rhinosinusitis. A controlled study.

1986 Aug 16

50 patients with chronic mucopurulent rhinosinusitis were randomly allocated to treatment with nasal sprays of dexamethasone, tramazoline, and neomycin, dexamethasone and tramazoline with no antibiotic, or matched placebo (propellant alone) four times daily to both nostrils for 2 weeks. The patients were assessed in a double-blind manner for symptomatic response and improvement in nasal mucociliary clearance, nasal airway resistance, sinus radiographs, and intranasal bacteriology and appearance. Both active preparations (with antibiotic 14 of 20 patients responded; without antibiotic 12 of 20 patients responded) were more effective than the placebo (2 of 10 patients responded). There was no significant difference in response between the active preparations with and without antibiotic. Thus, in treatment of chronic mucopurulent rhinosinusitis, reduction of the inflammatory response and decongestion make topical antibiotic unnecessary, probably by allowing host clearance mechanisms to recover.

Annotations:

Relevance

Population: Select

Intervention: Select

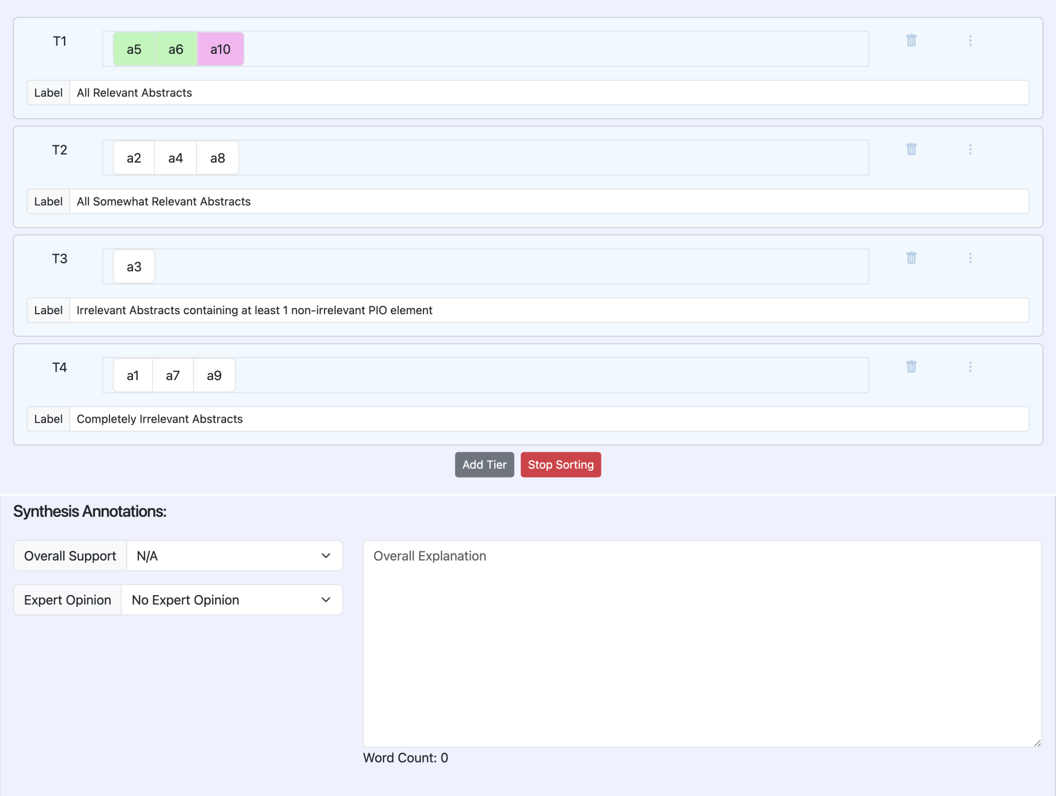
Outcome: Select

Overall: Select

Optional Comments

Comments

Figure 3: Presentation of claims, PIO elements, and abstracts in the annotation interface.



T1: a5 a6 a10

Label: All Relevant Abstracts

T2: a2 a4 a8

Label: All Somewhat Relevant Abstracts

T3: a3

Label: Irrelevant Abstracts containing at least 1 non-irrelevant PIO element

T4: a1 a7 a9

Label: Completely Irrelevant Abstracts

Synthesis Annotations:

Overall Support: N/A

Expert Opinion: No Expert Opinion

Overall Explanation

Word Count: 0

Figure 4: Presentation of the tiering and synthesis annotations interface.

dataset. We also present extracted PIO elements (See Appendix E) in a separate box with a rewritten claim created by inputting these elements in a template. All of this information was provided for the benefit of the annotator to clearly understand the claim in question.

Reading each abstract can be a cumbersome task. Therefore, we also provide for each abstract, information highlights and the published date of the paper associated with that abstract. These informational highlights covered the PIO elements in the abstract as well as the punchline of the abstract. This information is provided along with the abstracts in the TrialStreamer dataset.

Figure 4 shows the interface in which the expert annotators would tier abstracts and then provide the overall synthesis annotations with explanations. After the annotator is done with their relevance annotations, the interface will automatically tier abstracts according to these annotations. These automatic tiers are:

- **All Relevant Abstracts:** This tier contains all abstracts that were determined to be overall relevant to the claim.
- **All Somewhat Relevant Abstracts:** This tier contains all abstracts that were determined to be somewhat relevant to the claim.
- **Irrelevant Abstracts containing at least 1 non-irrelevant PIO element:** As the name suggests, this tier contains all irrelevant abstracts with at least 1 non-irrelevant PIO element in relation to the claim.
- **Compeletely Irrelevant Abstracts:** This tier contains all abstracts with the overall and all the PIO elements labeled as irrelevant in relation to the claim.

Expert annotators, when presented with these tiers, should try to further categories the collection of abstracts if possible. They are able to add tiers, manipulate their order, and change their names. They can also double click on an abstract tag, and the interface will display the abstract corresponding to that tag. All of these features serve to make the process of tiering abstracts as streamlined as possible for the expert annotators.