Medical Claim Checking: DREU Final Project Report

Lily Chen
MIT
l1ly@mit.edu

Abstract

Automated Claim Checking of medical claims made on social media can be extremely useful for clarifying critical assumptions about users' health and preventing the spread of harmful information. This paper presents a benchmark for an evidence-based approach to claim checking of medical claims with synthesis of randomized controlled trials (RCT) abstracts. This benchmark consists of 300 claims, with fine-grained evaluation and plain language explanations synthesizing medical evidence from experts.

We will publicly release our code after publication.

1 Introduction

Many users post statements or questions on social media about their health or medical knowledge, which often may include infactual parts. Without verification, this can spread misinformation that is dangerous and potentially deadly for people who utilize online information to inform their medical decisions. Our work attempts to develop and evaluate expert grade evidence-based systems to check medical claims by retrieving and synthesizing RCTs, a critical foundation of evidence based medicine that measure efficacy of new treatments (Sackett et al., 1996; Hariton and Locascio, 2018).

2 Medical Claim Checking Benchmark

The benchmark consists of expert claim checking assessments of 300 claims, each with 10 retrieved RCT abstracts as the corresponding medical evidence, and 300 expert written plain language explanations synthesizing the evidence to assess the claim.

2.1 Medical Claims from Social Media

The medical claims used in this benchmark are sourced from the RedHOT dataset (Wadhwa et al., 2023), which are directly extracted from various subreddits, and include annotation of Reddit health posts including claim extraction and PIO elements.

2.2 Retrieval

We used SOTA embedding model, `Alibaba-NLP/gte-large-en-v1.5`, which was the SOTA open-sourced embedding model with lightweight parameters, feasible to use to embed a database of 800,000 RCT abstracts with our limited resources. We embedded the concatenation of the RCT title and abstract content for the Trialstreamer database. We also embedded the post, claim, and its annotated PIO elements.

We retrieved the top 10 most relevant abstracts by cosine similarity.

2.3 Human Evaluation Framework

Relevance Population, Intervention, and Outcome elements are critical components of an RCT (Richardson et al., 1995).

Experts then evaluate the relevance of each of the PIO elements. Then, they rank the relevance of the abstract overall in respect to the claim.

If the Overall label of the abstract is graded as relevant, then the experts are asked to assess whether the support label of the abstract relative to the claim: (1) Supports, (2) Partially Supports, (3) Partially Refutes, and (4) Refutes.

Claim Support Positions

Tiering We introduce a tiering process to allow medical experts to analyze the quality of medical evidence. Our automatic tiers of abstracts are (1) Relevant, (2) Somewhat Relevant, (3) Irrelevant.

Synthesis Support and Expert Opinion Labels After completing the tiering step of the annotation and arranging a hierarchy of the quality of medical evidence in relation to the claim, annotators decide on a final support label of whether the body of evidence synthesized either (1) Supports, (2) Partially

Supports, (3) Inconclusive, (4) Partially Refutes, (5) Refutes the claim. This label is based on the evidence only.

Additionally, we include an optional label for experts to put their expert opinion from their clinical experience, which contain the same five support level options as the Synthesis Support label options. This allows us to clearly study the differences in expert opinion compared to the RCT evidence synthesis, as well as prevent the potential biases in expert evaluation of synthesis given their previous clinical experience.

Plain Language Explanations To justify their claim support label, experts write a paragraph length synthesis explanation. Our expert rationales followed the suggesting template:

- 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.

At the end, we include an opportunity for annotators to include a medical addendum of their medical expertise inputs on what is usually done in clinics in response to the claim, which could be very valuable to users.

2.4 Annotation

This benchmark was annotated by six medical experts, particularly one medical student in the U.S., four licensed doctors, and one Radiology Sciences researcher. They are all experienced with studying medical articles and synthesizing them to make critical decisions in biomedical research or treating patients.

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 to gauge whether they could follow instructions and gauge 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.

Following the recommendations of (Klie et al., 2024), the two co-first authors held a weekly group meeting with all medical experts to discuss the most disagreed examples, provide general feedback, and refine the annotation guidelines as needed.

Inter-Evaluator Agreement. Our inter-evaluator agreement for the pilot set of ten claims is shown in A.

3 Related Work

The recent surge in interest around fact-checking, particularly in scientific and health-related domains, has driven the creation of diverse datasets and models. These efforts, such as those focused on expert-curated questions (Malaviya et al., 2024), generating scientific claims (Wright et al., 2022), and improving claim verification with full-document context (Wadden et al., 2022b), highlight the need for robust verification across various contexts. Datasets like FakeCovid and COVIDLies (Shahi and Nandini, 2020; Hossain et al., 2020) emphasize the importance of cross-domain verification, while resources like SciFact-Open (Wadden et al., 2022a) address the complexities of open-domain scientific fact-checking.

In the health misinformation domain, datasets such as SCIFACT (Wadden et al., 2020), HealthVer (Sarrouti et al., 2021), and HealthFC (Vladika et al., 2024) provide frameworks for verifying health-related claims with evidence-based methods. These are complemented by broader resources like the Monant Medical Misinformation Dataset (Srba et al., 2022), which examines the mapping of medical misinformation, and SCITAB (Lu et al., 2023), which supports the compositional reasoning on scientific tables.

Expanding beyond health-specific contexts, general-purpose datasets and models such as FactKG (Kim et al., 2023), MISCCI (Glockner et al., 2024), and AVeriTeC (Schlichtkrull et al., 2023) underline the importance of reasoning and argumentation in fact-checking across diverse scenarios.

Additionally, research aligning social media medical claims with scientific evidence (Hughes

and Song, 2024) and new verification models based on causal graphs (Wu et al., 2023) contribute to the ongoing development of more comprehensive and effective fact-checking methodologies.

4 Future Work

There are many steps that remain to complete this project. First, we are working on improving the benchmark design by curating claims more carefully from disease populations that have more conducted RCTs. We are also looking into other retrieval ideas that could ensure higher quality of evidence retrieved. Additionally, we need to refine and finalize the annotation pipeline for medical experts, obtain better agreement, and scale up to the full 300 claims. Furthermore, on the automatic side, we need to complete many more baselines to explore the capabilities and shortcomings of automatic systems on this task. Plus, we need to do experiments focused on fine-tuning or synthetic data generation to contribute a tangible step forward to an automatic medical claim checking system.

Acknowledgments

I express my immense gratitude to Jessy Junyi Li for being an amazing DREU research advisor. Thank you also to the DREU program for this research opportunity and financial support.

References

Max Glockner, Yufang Hou, Preslav Nakov, and Iryna Gurevych. 2024. Missci: Reconstructing fallacies in misrepresented science. *arXiv preprint arXiv:2406.03181*.

Eduardo Hariton and Joseph J Locascio. 2018. Randomised controlled trials—the gold standard for effectiveness research. *BJOG: an international journal of obstetrics and gynaecology*, 125(13):1716.

Tamanna Hossain, Robert L. Logan IV, Arjuna Ugarte, Yoshitomo Matsubara, Sean Young, and Sameer Singh. 2020. COVIDLies: Detecting COVID-19 misinformation on social media. In *Proceedings of the 1st Workshop on NLP for COVID-19 (Part 2) at EMNLP 2020*, Online. Association for Computational Linguistics.

Anthony James Hughes and Xingyi Song. 2024. Identifying and aligning medical claims made on social media with medical evidence. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 8580–8593, Torino, Italia. ELRA and ICCL.

Jiho Kim, Sungjin Park, Yeonsu Kwon, Yohan Jo, James Thorne, and Edward Choi. 2023. FactKG: Fact verification via reasoning on knowledge graphs. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16190–16206, Toronto, Canada. Association for Computational Linguistics.

Jan-Christoph Klie, Richard Eckart de Castilho, and Iryna Gurevych. 2024. Analyzing Dataset Annotation Quality Management in the Wild. *Computational Linguistics*, pages 1–50.

Xinyuan Lu, Liangming Pan, Qian Liu, Preslav Nakov, and Min-Yen Kan. 2023. SCITAB: A challenging benchmark for compositional reasoning and claim verification on scientific tables. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7787–7813, Singapore. Association for Computational Linguistics.

Chaitanya Malaviya, Subin Lee, Sihao Chen, Elizabeth Sieber, Mark Yatskar, and Dan Roth. 2024. ExpertQA: Expert-curated questions and attributed answers. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 3025–3045, Mexico City, Mexico. Association for Computational Linguistics.

W S Richardson, M C Wilson, J Nishikawa, and R S Hayward. 1995. The well-built clinical question: a key to evidence-based decisions. *ACP J. Club*, 123(3):A12–3.

David L Sackett, William MC Rosenberg, JA Muir Gray, R Brian Haynes, and W Scott Richardson. 1996. Evidence based medicine. *BMJ: British Medical Journal*, 313(7050):170.

Mourad Sarrouti, Asma Ben Abacha, Yassine Mrabet, and Dina Demner-Fushman. 2021. Evidence-based fact-checking of health-related claims. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3499–3512, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Michael Schlichtkrull, Zhijiang Guo, and Andreas Vlachos. 2023. Averitec: A dataset for real-world claim verification with evidence from the web. In *Advances in Neural Information Processing Systems*, volume 36, pages 65128–65167. Curran Associates, Inc.

Gautam Kishore Shahi and Durgesh Nandini. 2020. *FakeCovid- A Multilingual Cross-domain Fact Check News Dataset for COVID-19*. ICWSM.

Ivan Srba, Branislav Pecher, Matus Tomlein, Robert Moro, Elena Stefancova, Jakub Simko, and Maria Bielikova. 2022. Monant medical misinformation dataset: Mapping articles to fact-checked claims. In *Proceedings of the 45th International ACM SIGIR*

Conference on Research and Development in Information Retrieval, SIGIR '22, page 2949–2959, New York, NY, USA. Association for Computing Machinery.

Juraj Vladika, Phillip Schneider, and Florian Matthes. 2024. HealthFC: Verifying health claims with evidence-based medical fact-checking. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 8095–8107, Torino, Italia. ELRA and ICCL.

David Wadden, Shanchuan Lin, Kyle Lo, Lucy Lu Wang, Madeleine van Zuylen, Arman Cohan, and Hannaneh Hajishirzi. 2020. Fact or fiction: Verifying scientific claims. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7534–7550, Online. Association for Computational Linguistics.

David Wadden, Kyle Lo, Bailey Kuehl, Arman Cohan, Iz Beltagy, Lucy Lu Wang, and Hannaneh Hajishirzi. 2022a. SciFact-open: Towards open-domain scientific claim verification. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 4719–4734, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

David Wadden, Kyle Lo, Lucy Lu Wang, Arman Cohan, Iz Beltagy, and Hannaneh Hajishirzi. 2022b. MultiVerS: Improving scientific claim verification with weak supervision and full-document context. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 61–76, Seattle, United States. Association for Computational Linguistics.

Somin Wadhwa, Vivek Khetan, Silvio Amir, and Byron Wallace. 2023. RedHOT: A corpus of annotated medical questions, experiences, and claims on social media. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 809–827, Dubrovnik, Croatia. Association for Computational Linguistics.

Dustin Wright, David Wadden, Kyle Lo, Bailey Kuehl, Arman Cohan, Isabelle Augenstein, and Lucy Lu Wang. 2022. Generating scientific claims for zero-shot scientific fact checking. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2448–2460, Dublin, Ireland. Association for Computational Linguistics.

Jinxuan Wu, Wenhan Chao, Xian Zhou, and Zhunchen Luo. 2023. Characterizing and verifying scientific claims: Qualitative causal structure is all you need. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 13428–13439, Singapore. Association for Computational Linguistics.

A Pilot Inter-evaluator Agreement

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

Table 1: Inter-evaluator agreement measured through Randolph's κ for the training set of 10 claims.