# "Devils Are in the Details": Annotating Specificity of Clinical Advice from Medical Literature

**Yingya Li**
School of Information Studies
Syracuse University
yli48@syr.edu

**Bei Yu**
School of Information Studies
Syracuse University
byu@syr.edu

## Abstract

Prior studies have raised concerns over specificity issues in clinical advice. Lacking specificity — explicitly discussed detailed information — may affect the quality and implementation of clinical advice in medical practice. In this study, we developed and validated a fine-grained annotation schema to describe different aspects of specificity in clinical advice extracted from medical research literature. We also presented our initial annotation effort and discussed future directions towards an NLP-based specificity analysis tool for summarizing and verifying the details in clinical advice.

## 1 Introduction

In medical literature, authors often explain clinical implications after presenting their research findings. For example, "Results of this post-hoc analysis suggest that LEV may be a suitable option for initial monotherapy for patients aged $\geq$ 60 years with newly diagnosed epilepsy" (Pohlmann-Eden et al., 2016). Clinical advice like this can influence health researchers and practitioners on specific medical practices. Hence, it is an important information service to retrieve and analyze clinical advice from medical literature.

Prior studies have identified some quality issues that may affect the implementation of clinical advice. One concern is the lack of specificity. Two studies have compared the implementation outcomes of professionally-designed clinical guidelines with different levels of specificity, and found that concrete and precise descriptions resulted in higher adoption rate (Michie and Johnston, 2004; Michie and Lester, 2005). Clinical advice in medical literature also varies in specificity levels. For example, advice sentences appeared in abstracts tend to be less specific compared to those in discussions, where more space is available for explaining the details (Li and Yu, 2022).

To better retrieve and summarize clinical advice from medical literature, this study aims to develop a taxonomy of specifics in clinical advice, such that they may be retrieved and compared in finer granularity. We developed and validated an annotation schema that can partition a clinical advice sentence to multiple elements: 1) agents; 2) substantial qualifications or elaborations; 3) chain of reasoning; 4) confidence. This annotation schema was developed based on medical research on clinical guidelines and NLP research on modeling specificity as a language construct.

We also discussed the future directions for computationally modeling specificity of clinical advice in medical literature. Such specificity analysis tool can be used for downstream applications such as detecting quality issues of clinical advice. For example, one study raised severe concern that many recommendations for clinical practice were not supported by findings in the conclusions (Yavchitz et al., 2016). The problem is more severe in abstracts than in discussions. Since abstracts are much more accessible than full-text articles, the "spins" in abstracts are also more harmful than those in discussions (Boutron et al., 2014). An NLP-based specificity analysis tool can help compare recommendation details against available evidence, or compare similar recommendations in fine-granularity.

## 2 Related Work

Specificity is an important concept in both clinical practice and claim analyses. In medical domain, specificity is defined narrowly, focusing on the detailed information regarding clinical practice and health-related behavior changes. For example, Shekelle et al. (2000) defined a specific guideline as "creates clinical appropriate criteria for a large number of clinically detailed patient presentations; it does not force consensus" (p.1431). Similarly, Michie and Lester (2005) argued that specific



*Proceedings of the Second Workshop on Understanding Implicit and Underspecified Language*, pages 17 - 21

clinical guidelines give "detailed advice on which performance is appropriate in which situation and in what patient group and determining which factors, or conditions should be taken into account" (p. 367). Note that clinical guidelines used in practice are usually developed by professional institutions such as National Institute for Clinical Excellence (NICE). They are usually more comprehensive and specific than the clinical advice from individual research papers.

Compared to the narrow definition in clinical domain, specificity is defined more broadly in the NLP field, referring to how much detailed information is included in a statement. Depending on the text domains, researchers have proposed different taxonomies to define specificity. For example, in education domain, the specificity of classroom discussions was defined based on four aspects: "involves one character or scene", "gives substantial qualifications or elaboration", "uses content-specific vocabulary", and "provides a chain of reasoning" (Lugini and Litman, 2017). Similarly, arguments in student essays were assigned specificity scores based on occurrence of qualifiers, references to supporting components, hypotheses, and real-world examples (Carlile et al., 2018). Specificity in other domains was defined quite differently. For example, the specificity of pledges of election manifestos were labelled based on expressions of moral values, intangible goals and outcomes, commitment to the maintenance of functioning policy, means and details to achieve the objectives (Subramanian et al., 2019). The specificity in social media posts was defined based on their references to specific person, object or event (Gao et al., 2019).

Although the exact aspects applied to describe specificity differ by domains, they usually cover the anwers to questions about who, what, when, where, why, and how. Since the clinical domain and the education domain are most relevant to our task, we defined our annotation schema by combining the definitions from these two domains.

## 3 Data and Annotation Schema

### 3.1 Dataset

In this study, we used an open-access dataset on health advice, which contains a sample of 10,848 sentences extracted from abstracts and discussion sections in medical research papers, in which 2,748 sentences were annotated as health advice (Li et al., 2021). The research papers include different study designs, including randomized controlled trials and four types of observational studies, including cross-sectional, case-control, retrospective, and prospective studies. We sampled sentences from all study designs to ensure the annotation schema is generalizable. We first sampled 100 advice sentences to develop the annotation schema and finalize the definition of each concept. We then sampled another 100 advice sentences to evaluate the inter-coder agreement on the finalized annotation schema.

### 3.2 Annotating Clinical and Non-clinical Advice

In the health advice dataset (Li et al., 2021), the annotated health advice may recommend clinical intervention and practice ("clinical advice") or simply raise awareness and call for actions for certain health behavior or policy change ("non-clinical advice"). The latter type tends to use vague verbs such as "address" and "encourage" instead of concrete description of interventions. In this study, we focus on clinical advice. Hence, the first step in the annotation is to distinguish clinical vs. non-clinical advice. Clinical advice will be further annotated with specificity aspects. Occasionally, we encountered a sentence with serious semantic ambiguity, and labelled it as incomprehensive.

Drawing on prior specificity annotations on clinical guidelines, we adopted two key aspects that also appear in clinical advice in medical literature: "agents" and "substantial qualifications or elaborations". In addition, we found two aspects in clinical advice from medical literature but are absent in clinical guidelines: "chain of reasoning" and "confidence". Diffrent from clinical guidelines that focus on what to do only, clinical advice from research papers sometimes includes explanations on the reason of why a recommendation was made. Therefore, we added "chain of reasoning". This concept is borrowed from specificity annotation in the education domain (Lugini and Litman, 2017).

In addition, authors often expressed their confidence in clinical advice using words like "possible", "may","can", and "is", based on the evidence level. This concept is relevant to the "strong/weak advice" concept in the original health advice data set, or prior studies that distinguished "implicit/explicit advice" (Sumner et al., 2014). These prior studies aimed for categorical definition of the advice strength, and they cover both clinical and non-clinical advice. In this study, we use the concept

| Advice Type | Description | Example Sentence and Specificity Annotation |
|---|---|---|
| Non-clinical Advice | Health advice that aims to raise awareness or calls for actions for health-related behavioral changes. The outcome of the action is not directly measurable. Use verbs such as "address", "encourage", and "ensure". | 1. Special attention is required in such patients while doing treatment planning.<br>2. We conclude that it is important to encourage physical activity in this population. |
| Clinical Advice | Health advice that provides clear actionable suggestions for medical practice and policy changes. The advice contains precise and concrete description for the treatment or intervention that needs to be taken. | 3. Therefore, intraoperative antifibrinolysis may not be indicated in routine cardiac surgery when other blood-saving techniques are adopted.<br>**Annotation**: agent (N/A), intervention ("intraoperative antifibrinolysis"), target (N/A), goal ("routine cardiac surgery when other blood-saving techniques are adopted"), chain of reasoning (N/A), confidence ("may not be indicated in")<br>4. Therefore, due to the cost, possible side effects, and the limited saving of homologous blood, intraoperative antifibrinolytic therapy may not be indicated in routine cardiac surgery.<br>**Annotation**: agent (N/A), intervention ("intraoperative antifibrinolytic therapy"), target (N/A), goal ("routine cardiac surgery"), chain of reasoning ("therefore, due to the cost, possible side effects, and the limited saving of homologous blood"), confidence ("may not be indicated in") |

Table 1: Specificity annotation schema and sentence examples.

| | RCTs | Cross-Sectional | Case-Control | Retrospective | Prospective | Total | Percentage |
|---|---|---|---|---|---|---|---|
| Clinical | 27 | 15 | 17 | 22 | 19 | 100 | 50.0% |
| Non-clinical | 13 | 25 | 22 | 18 | 20 | 98 | 49.0% |
| Total | 40 | 40 | 39 | 40 | 39 | 200 | |

Table 2: Distribution of clinical and non-clinical advice in annotated corpus.

"confidence" to emphasize that we aim to identify the phrases that describe confidence level in clinical advice only.

Overall, we defined specificity from the following four dimensions: "agents", "substantial qualifications or elaborations", "chain of reasoning", and "confidence". Table 1 shows the definition and sentence examples of the annotation schema.

**Agents**: the party to carry out the recommended clinical practice, such as health practitioners or organizations.

**Substantial qualifications or elaborations**: concrete and precise details in health advice that depicts what, who, when, where, and how information to assist implementation of actionable clinical practice. We further categorized it by the following sub-dimensions:

*Intervention*: the details of treatment, such as therapy procedures, doses and usage

*Target*: the party to receive the recommended intervention, usually patients, sometimes including demographical details or body parts to be treated.

*Goal*: illness/symptom that the intervention aims to treat, or another treatment that it aims to support.

**Chain of reasoning**: reasons for the clinical advice, normally indicated by linguistic cues such as "although", "as long as" and "since", when health researchers admitting a fact or showing contrasts in recommendations.

**Confidence**: the level of confidence researchers have when giving the advice.

### 3.3 Inter-coder Agreement

To test the validity of the proposed schema, a sample of 100 advice sentences were randomly selected for inter-coder agreement evaluation. We applied disproportionate stratified sampling to get 20 advice sentences from each of the 5 study designs. Two annotators with the education backgrounds of linguistics and information science each labelled the 100 sentences for clinical advice and specificity. The overall Cohen's Kappa agreement (Cohen, 1960) on annotating clinical and non-clinical advice was 0.88, indicating a near-perfect inter-coder agreement (McHugh, 2012). The agreement on each of the specificity dimensions were: agent (0.98), intervention (0.93), target (0.91), goal (0.87), chain of reasoning (0.91), and confidence

| Specifics | Count | Percentage |
|---|---|---|
| Agent | 3 | 3.0% |
| Intervention | 100 | 100.0% |
| Target | 58 | 58.0% |
| Goal | 88 | 88.0% |
| Reasoning | 25 | 25.0% |
| Confidence | 100 | 100.0% |

Table 3: Distribution of the advice details on each dimension of specificity.

(0.93). Disagreed cases were later resolved by the two annotators through discussion.

### 3.4 Specifics in Clinical Advice

We annotated 200 health advice sentences in total for schema development and validation. Excluding two incomprehensible sentences, 100 were "clinical advice", and 98 were "non-clinical" advice. Table 2 shows their distributions across different study designs. The almost equal distribution of clinical and non-clinical advice suggests that researchers tend to give both advice for clinical practice/interventions and advice that calls for general health-related behavior changes. However, when zooming into the different study deigns, we noted that RCTs have a higher percentage of clinical advice (67.5%) than the observational studies followed by retrospective (55.0%), prospective (47.5%), case-control (42.5%), and cross-sectional studies (37.5%), indicating that researchers are more likely to give clinical advice in studies with higher evidence levels. The quality of clinical advice given in observational studies was more often questioned by the research community (Cofield et al., 2010).

Among the 100 clinical advice sentences, "intervention", "confidence" and "goal" are most often mentioned. Different from professionally-designed clinical guidelines, "agent" in medical literature is almost always omitted, and "target" is omitted over 40% of times. Reasoning is also not often provided (25%). See Table 3 for the aspect distribution.

With the fine-grained specificity annotation, we can then compare details of recommendations against evidence strength or compare different versions of similar recommendations. For example, in Table 1, examples 3 and 4 appear in the same research paper but different sections. The annotations show that the first sentence provides a more specific goal, while the second sentence provides reasoning.

## 4 Towards Computational Modeling of Specificity

The explosive growth of research output and restricted human capacities in information processing and decision making calls for an NLP-based specificity analysis tool to synthesize and aggregate the scientific evidence and clinical recommendations in research publications. The developed annotation schema may then be used to develop automatic prediction models for clinical advice specifics classification and specifics extraction. Based on the occurrence for each specificity dimension, we could frame the task as a sentence-level classification task and to computationally model the specificity level in each advice sentences. Utilizing the annotated details under each specificity aspect, the task could also be framed as an information extraction one. Information extraction tools could be developed to extract the details of each recommendation. For example, simple rule-based approaches using regular expressions (Savova et al., 2010a) may identify the aspects of agents and targets. Existing NLP tools for medical concepts (e.g. Savova et al., 2010b; Zhou et al., 2019; Zhang et al., 2021) and clinical relation extractions based on pre-trained language models such as BERT (Roy and Pan, 2021) may further identify other specificity aspects in the develop schema. After extracting the specifics explicitly mentioned in each recommendation, we could compare different versions of semantically similar recommendations across the specifics to detect the inconsistent or exaggerated clinical advice in research literature.

## 5 Conclusion

In this work we presented a fine-grained annotation schema for describing specificity in clinical advice extracted from medical research literature. An inter-coder agreement check shows the proposed annotation schema reached almost perfect agreement in all dimensions. The annotation schema could be used to develop gold-standard dataset that can be used to develop NLP models for identifying fine-grained specificity aspects in clinical advice, and to support downstream applications such as summarizing clinical advice or fact checking.

## Acknowledgement

This research is supported by the US National Science Foundation under grant 1952353 and the Syracuse University CUSE Grant.

## References


Isabelle Boutron, Douglas G Altman, Sally Hopewell, Francisco Vera-Badillo, Ian Tannock, and Philippe Ravaud. 2014. Impact of spin in the abstracts of articles reporting results of randomized controlled trials in the field of cancer: the spiin randomized controlled trial. *Journal of Clinical Oncology*, 32(36):4120–4126.

Winston Carlile, Nishant Gurrapadi, Zixuan Ke, and Vincent Ng. 2018. Give me more feedback: Annotating argument persuasiveness and related attributes in student essays. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 621–631.

Stacey S Cofield, Rachel V Corona, and David B Allison. 2010. Use of causal language in observational studies of obesity and nutrition. *Obesity facts*, 3(6):353–356.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46.

Yifan Gao, Yang Zhong, Daniel Preoţiuc-Pietro, and Junyi Jessy Li. 2019. Predicting and analyzing language specificity in social media posts. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 6415–6422.

Yingya Li, Jun Wang, and Bei Yu. 2021. Detecting health advice in medical research literature. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6018–6029.

Yingya Li and Bei Yu. 2022. Advice giving in medical research literature. In *International Conference on Information*, pages 261–272. Springer.

Luca Lugini and Diane Litman. 2017. Predicting specificity in classroom discussion. In *Proceedings of the 12th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 52–61.

Mary L McHugh. 2012. Interrater reliability: the kappa statistic. *Biochemia medica*, 22(3):276–282.

S Michie and Kathryn Lester. 2005. Words matter: increasing the implementation of clinical guidelines. *BMJ Quality & Safety*, 14(5):367–370.

Susan Michie and Marie Johnston. 2004. Changing clinical behaviour by making guidelines specific. *Bmj*, 328(7435):343–345.

Bernd Pohlmann-Eden, Anthony G Marson, Matthias Noack-Rink, Francisco Ramirez, Azita Tofighy, Konrad J Werhahn, Imane Wild, and Eugen Trinka. 2016. Comparative effectiveness of levetiracetam, valproate and carbamazepine among elderly patients with newly diagnosed epilepsy: subgroup analysis of the randomized, unblinded komet study. *BMC neurology*, 16(1):1–12.

Arpita Roy and Shimei Pan. 2021. Incorporating medical knowledge in bert for clinical relation extraction. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5357–5366.

Guergana K Savova, Jin Fan, Zi Ye, Sean P Murphy, Jiaping Zheng, Christopher G Chute, and Iftikhar J Kullo. 2010a. Discovering peripheral arterial disease cases from radiology notes using natural language processing. In *AMIA Annual Symposium Proceedings*, volume 2010, page 722. American Medical Informatics Association.

Guergana K Savova, James J Masanz, Philip V Ogren, Jiaping Zheng, Sunghwan Sohn, Karin C Kipper-Schuler, and Christopher G Chute. 2010b. Mayo clinical text analysis and knowledge extraction system (ctakes): architecture, component evaluation and applications. *Journal of the American Medical Informatics Association*, 17(5):507–513.

Paul G Shekelle, Richard L Kravitz, Jennifer Beart, Michael Marger, Mingming Wang, and Martin Lee. 2000. Are nonspecific practice guidelines potentially harmful? a randomized comparison of the effect of nonspecific versus specific guidelines on physician decision making. *Health services research*, 34(7):1429.

Shivashankar Subramanian, Trevor Cohn, and Timothy Baldwin. 2019. Deep ordinal regression for pledge specificity prediction. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1729–1740.

Petroc Sumner, Solveiga Vivian-Griffiths, Jacky Boivin, Andy Williams, Christos A Venetis, Aimée Davies, Jack Ogden, Leanne Whelan, Bethan Hughes, Bethan Dalton, et al. 2014. The association between exaggeration in health related science news and academic press releases: retrospective observational study. *Bmj*, 349.

Amélie Yavchitz, Philippe Ravaud, Douglas G Altman, David Moher, Asbjørn Hrobjartsson, Toby Lasserson, and Isabelle Boutron. 2016. A new classification of spin in systematic reviews and meta-analyses was developed and ranked according to the severity. *Journal of clinical epidemiology*, 75:56–65.

Yuhao Zhang, Yuhui Zhang, Peng Qi, Christopher D Manning, and Curtis P Langlotz. 2021. Biomedical and clinical english model packages for the stanza python nlp library. *Journal of the American Medical Informatics Association*, 28(9):1892–1899.

Xin Zhou, Yanshan Wang, Sunghwan Sohn, Terry M Therneau, Hongfang Liu, and David S Knopman. 2019. Automatic extraction and assessment of lifestyle exposures for alzheimer's disease using natural language processing. *International journal of medical informatics*, 130:103943.