Summer 7-1-2022

An NLP Analysis of Health Advice Giving in the Medical Research Literature

Yingya Li
Syracuse University

Recommended Citation

Li, Yingya, "An NLP Analysis of Health Advice Giving in the Medical Research Literature" (2022). *Dissertations - ALL*. 1564.
https://surface.syr.edu/etd/1564

Abstract

Health advice – clinical and policy recommendations – plays a vital role in guiding medical practices and public health policies. Whether or not authors should give health advice in medical research publications is a controversial issue. The proponents of "actionable research" advocate for the more efficient and effective transmission of science evidence into practice. The opponents are concerned about the quality of health advice in individual research papers, especially that in observational studies. Arguments both for and against giving advice in individual studies indicate a strong need for identifying and accessing health advice, for either practical use or quality evaluation purposes. However, current information services do not support the direct retrieval of health advice. Compared to other natural language processing (NLP) applications, health advice has not been computationally modeled as a language construct either. A new information service for directly accessing health advice should be able to reduce information barriers and to provide external assessment in science communication.

This dissertation work built an annotated corpus of scientific claims that distinguishes health advice according to its occurrence and strength. The study developed NLP-based prediction models to identify health advice in the PubMed literature. Using the annotated corpus and prediction models, the study answered research questions regarding the practice of advice giving in medical research literature. To test and demonstrate the potential use of the prediction model, it was used to retrieve health advice regarding the use of hydroxychloroquine (HCQ) as a treatment for COVID-19 from LitCovid, a large COVID-19 research literature database curated by the National Institutes of Health.

An evaluation of sentences extracted from both abstracts and discussions showed that BERT-based pretrained language models performed well at detecting health advice. The health

advice prediction model may be combined with existing health information service systems to provide more convenient navigation of a large volume of health literature. Findings from the study also show researchers are careful not to give advice solely in abstracts. They also tend to give weaker and non-specific advice in abstracts than in discussions. In addition, the study found that health advice has appeared consistently in the abstracts of observational studies over the past 25 years. In the sample, 41.2% of the studies offered health advice in their conclusions, which is lower than earlier estimations based on analyses of much smaller samples processed manually. In the abstracts of observational studies, journals with a lower impact are more likely to give health advice than those with a higher impact, suggesting the significance of the role of journals as gatekeepers of science communication.

For the communities of natural language processing, information science, and public health, this work advances knowledge of the automated recognition of health advice in scientific literature. The corpus and code developed for the study have been made publicly available to facilitate future efforts in health advice retrieval and analysis. Furthermore, this study discusses the ways in which researchers give health advice in medical research articles, knowledge of which could be an essential step towards curbing potential exaggeration in the current global science communication. It also contributes to ongoing discussions of the integrity of scientific output.

This study calls for caution in advice-giving in medical research literature, especially in abstracts alone. It also calls for open access to medical research publications, so that health researchers and practitioners can fully review the advice in scientific outputs and its implications. More evaluative strategies that can increase the overall quality of health advice in research articles are needed by journal editors and reviewers, given their gatekeeping role in science communication.

AN NLP ANALYSIS OF HEALTH ADVICE GIVING IN THE MEDICAL RESEARCH LITERATURE

by

Yingya Li

B.A., University of Science and Technology Beijing, 2013
M.A., Syracuse University, 2015

Dissertation
Submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy in
Information Science and Technology

Syracuse University

July 2022

To my parents

Acknowledgement

I would like to thank my advisor, Dr. Bei Yu for her generous guidance and support over the past six and a half years. Dr. Yu inspired my interest in the field of information science and natural language processing. With great patience, she taught me from scratch how to do research. It has been a great pleasure and rewarding experience to learn from her and work with her. I have always been inspired by her wisdom and rigor of research and her generosity and kindness. I feel very fortunate to be her student.

I am grateful to Dr. Jun Wang, whom I worked with closely during my doctoral training. I appreciate his solid help in research and machine learning techniques. I would like to thank my dissertation committee members Drs. Jian Qin, Daniel Acuna, and Jeffrey Stanton for their helpful comments and advice for the thesis. Special thanks to Dr. Jodi Schneider from UIUC and Dr. Yiqi Li for kindly serving as the external and internal readers of my dissertation, and to Dr. Dessa Bergen-Cico from the Department of Public Health for chairing my dissertation defense.

I would also like to thank Drs. Caroline Haythornthwaite, Steve Sawyer, Jennifer Stromer-Galley, Kevin Crowston, Jason Dedrick and many of the faculty members at the iSchool of Syracuse University for their help and advice. Thanks to Dr. Amanda Stent, who took me as an intern in the summer of 2021 at Bloomberg and provided me with valuable suggestions about research and career development.

Thanks to my fellow graduate students and friends for all their help: Qunfang Wu, Mahboobeh Harandi, Sikana Tanupabrungsun, Yisi Sang, Han Zhuang, Feifei Zhang, Alain Shema, Yaxing Yao, Jun Zhang, Qinyun Zhu, Mahmuda Rahma, Li Chen, Chunxu Tang, Chengxiang Yin, and many others. I hope that I was also able to help them as much as they have helped me.

Finally, thanks to my parents for their unconditional love and support. This thesis is dedicated to them.

Preliminary results of this dissertation were published as the following:

[1] Li, Y., Wang, J., & Yu, B. (2021, November). Detecting Health Advice in Medical Research Literature. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing* (pp. 6018–6029).

[2] Li, Y., & Yu, B. (2022, February). Advice Giving in Medical Research Literature. In *International Conference on Information* (pp. 261–272). Springer, Cham.

Table of Contents

List of Figures

List of Tables

Chapter 1 Introduction

1.1 Background and Motivation

Evidence-based health advice from scientific literature guides medical practice and public health policies. However, whether or not to give health advice based on the results of a single study is a controversial issue. The proponents of “actionable research” would like to encourage the more efficient and effective transformation of science evidence into practice (Green et al., 2009). If researchers themselves do not discuss the practical value of their findings, press officers and journalists might misinterpret the results and give exaggerated health advice in press releases and news articles (Sumner et al., 2014; Haneef et al., 2015). Opponents argue that clinical and policy recommendations on health-related issues should not be allowed in research papers. This is because a single paper may lack sufficient information about all the evidence in real practice and there is limited manuscript space for a full review of alternative choices (Cummings, 2007).

The quality of health advice in observational studies is of particular concern to scientists and researchers. Both observational studies and randomized controlled trials (RCTs) are common research designs used in the health and medical domains. In evidence-based medicine (EBM) (Sackeit et al., 1996), RCTs aim to establish causal relationships, while observational studies aim to confirming associations between exposures and outcomes. The evidence from observational studies may be the best available when RCTs are impossible or unethical (Song and Chung 2010). However, the potential overinterpretation of observational results has led to arguments against the value of health advice derived from observational studies in guiding health decisions (Banerjee and Prasad 2020). Some medical experts warn that a large proportion of such health advice is not fully supported by the studies associated with it (Wilson and Chestnutt, 2016; Banerjee and Prasad, 2020).

Although a journal's impact factor can be an indicator of quality for general medical journals (Saha et al., 2003), health advice based on overinterpreted observational results is found to be common in leading medical journals (Prasad et al., 2013). Previous manually conducted content analyses of science communication also found opposing results regarding the advice given in journals with high and low impacts. For example, Lumbreras et al. (2009) found that articles published in journals with higher impact factors were more likely to over-interpret their findings for clinical applicability than those with lower impact factors. On the contrary, Wilson and Chestnutt (2016) noted that observational studies in journals with lower impact factors were more likely to have clinical and policy recommendations compared to those published in journals with a high impact. However, these prior studies were based on small-scale content analyses of certain specific health topics. Whether the findings are generalizable to the entire observational study literature, regardless of the health topic, remains an open question.

Besides the question of whether to give health advice based on individual study results, medical researchers also face the question of where to give health advice. In practice, researchers can choose to give advice in the abstract section, or in the discussion section after presenting the study results, or both. Prior studies have questioned the quality of health advice in abstracts, many of which misinterpret the research findings, claim exaggerated significance, or give inadequate recommendations for practice in certain clinical areas (Lazarus et al., 2015; Cooper et al., 2019). In addition,, abstracts offer very little room for giving detailed advice. At the same time, they are especially accessible to the public, and thus can have a broad audience. The discussion sections of articles, which provide more room to discuss implications, are often part of full-text offerings behind paywalls, although they allow for more room to discuss the implications (Hopewell et al., 2008). Little attention has been given to examine whether the recommendations given in abstracts and discussions are equivalent in the amount of information they provide. If major differences exist, then the paywalls are a barrier to accessing the complete information; otherwise, a health advice retrieval system just needs to retrieve advice in abstracts.

The debate over whether and how to give health advice in individual studies indicates a strong need for identifying and accessing health advice, for either practical use or quality evaluation purposes. However, navigating the large volume of medical papers is a daunting task (Straus and Haynes, 2009; Fry and Attawet, 2018), and outdated information systems have impeded access to health advice (Green et al., 2009). Also, the fast growth of the medical and health literature further exacerbates the challenge (Williamson and Minter, 2019). For example, the most recent COVID-19 outbreak has brought an explosion of research output about the disease (Brainard, 2020). While scientists around the world are racing to understand the transmission, prevention, and treatment of the disease, the fast-changing evidence has been challenging medical experts, governments, and the public in their quest to make informed decisions.

The strong need for understanding the fast-growing scientific evidence in COVID-19 has led to the creation of specialized data hubs and search platforms. NIH has created LitCovid, a curated literature hub for tracking up-to-date scientific information about the disease (Chen, Allot, and Lu, 2020). Allen AI has partnered with researchers to release CORD-19, a free source of more than 130,000 scholarly articles about the novel coronavirus (Wang et al., 2020). Several literature search and visualization systems powered by machine learning and NLP techniques, such as COVID-19 Navigators (IBM Watson, 2020), SciSight (Hope et al., 2020), and COVID Scholar (UC Berkeley, 2020) have also been developed. Nevertheless, current data hubs and information services do not support the direct retrieval of health advice unless one acquires access to the full-text content of research papers. Researchers and health practitioners still need to spend a lot of time gathering supporting and opposing evidence, for example, on whether hydroxychloroquine (HCQ) is a viable COVID-19 treatment. HCQ was considered a promising treatment option at the beginning of the COVID-19 pandemic (Gautret et al., 2020); however, early studies that recommended using HCQ were later criticized for a lack of randomization in their study designs. The resulting conflicting evidence fueled high-stakes debates on news and social media about the efficacy of HCQ (Pillar, 2020). In later clinical trials, with new evidence and randomized study results, HCQ was found to be ineffective (Lewis et al., 2021). Based on ongoing analysis and emerging scientific data, the U.S. Food and Drug Administration revoked its emergency-use authorization to treat certain hospitalized COVID patients with HCQ and chloroquine, as these medicines showed no benefits for decreasing the likelihood of death or speeding recovery. To make sense of situations like this without the support of a direct retrieval of advice, researchers and practitioners still need to spend a lot of time gathering the conflicting information and evolving advice for combating COVID-19.

Driving the need for automatic extraction and concerns about the quality concern of health advice, is the need to comprehensively understand the status of extrapolating health advice in individual papers (abstracts vs. discussions), and among different journals (high-impact vs. low-impact) over the years. Though many efforts have been made to identify health advice, most studies have been done on a small scale, using the manual analysis approach (e.g., Prasad et al., 2013; Sumner et al., 2014; Wilson and Chestnutt, 2016). These studies were useful for establishing a snapshot view in the past. However, the significant time and labor costs required for manual analysis create not only a need for a feasible machine-learning and NLP-based computational approach for detecting advice in large amounts of research publications, news articles, and online posts, but also for ways to track the creation and diffusion of health advice through domains and over time.

Prior linguistic studies have provided rich theoretical taxonomies of language phenomena related to health advice, while NLP techniques enable us to automate the identification of these language phenomena. Health advice as a language construct is closely linked to the linguistic concept of imperative language, which conveys a speaker's demand for action (Condoravdi and Lauer, 2012). In the case of advice, language indicators such as hedges, modalities, evidentials can show a speaker's level of commitment and indicate the strength of the advice. However, most work in the linguistic studies is based on small sample sizes. Without accommodating the rules associated with language use in different contexts or automating the process of advice identification, researchers using observation may arrive at norms applicable only to restricted situations. Though a variety of computational methods have been developed to extract health-related expressions, patterns, and components in different contexts (e.g., Light et al., 2004; Kwong and Yorke-Smith, 2009; Wei et al., 2013; Mao et al., 2014; Qian et al., 2016), similar tasks in the NLP domain, such as suggestion mining (Negi, Daudert, and Buitelaar, 2019), have not been well explored. We are still lacking in generalized definitions about health advice and systematic methods for measuring it. Moreover, ways to apply language technologies to automatically detect health advice and its level of commitment have not been adequately researched.

1.2 Research Goal

This dissertation work aims to answer important research questions about the automatic recognition of health advice and the practice of offering health advice in medical research literature. The computational approaches to be developed here can be used for external assessments of the quality of health advice in research publications and can help the public to judge the validity and value of health advice when making medical decisions.

This study sought empirical answers to the following research questions:

RQ1: To what extent can NLP prediction algorithms detect health advice in PubMed publications?

RQ2: Where do research papers give health advice? If a research paper gives health advice in both the abstract and the discussion, are the advice statements equivalent?

RQ3: Is health advice prevalent in observational studies? How have patterns changed over time?

RQ4: Do journals differ in their practice of allowing advice giving or not?

RQ5. What health advice has been offered regarding the use of HCQ for treating COVID-19?

1.3 Significance of the Study

This dissertation work proposed an NLP-based approach for automatically extracting health advice from medical research papers and analyzing it. As the first of its kind, the research will contribute to the fields of information science, NLP, and public health by providing a new prediction model for identifying health advice in medical literature and by providing new evidence from large-scale analysis for answering research questions regarding the status of health advice in scientific publications. Specifically, the methods and findings from this research make the following contributions:

- The resulting annotation taxonomy and corpus of health advice will serve as valuable resources for mining the patterns and trends in the giving of health advice, which can have a significant impact on science communication and education.
- The research will broaden our understanding of health advice as a language construct, which will foster an understanding of the language that different information stakeholders use when giving health advice.

- The research will advance our knowledge about the automated recognition of health advice in scientific literature.
- The findings will also expand our understanding of the practice in the medical research literature of giving health advice, which could be an essential step towards curbing the exaggeration of health claims in the current global science communication.

Practically, this research is expected to broadly benefit society in the following ways:

- Given the increasing impact of scientific discoveries on people's everyday lives and on public health policies, computational approaches to detecting health advice will help researchers keep track of the implications of the most recent studies in scientific publications. It will also help them monitor the validity of scientific research and ensure the availability of reliable evidence for supporting individual and government decision making.
- The prediction model for health advice identification can service as the core function of an information service for analyzing health advice in scientific publications. Such a health advice service could be integrated with existing data hubs, and this could help answer important research questions, such as "What health advice has been given regarding the use of certain medicines (e.g., HCQ or remdesivir) for COVID-19 treatment?" If the output provided by the model were to be combined with other metadata on publication venues (e.g., journal rankings) and study designs (e.g., RCTs or observational studies), the health advice service would be able to organize health advice based on the strength of evidence. If it were to be combined with other NLP tools (e.g., stance classification or sentiment analysis), the service would be able to compare the evidential strength of recommendations for or against certain treatments.

- The prediction model could also be used to analyze health advice that appears in research news and institutional press releases and on social media platforms. By comparing the advice with the actual scientific findings and implications, it could track the inaccuracies and to monitor the quality of scientific communications for the general public.

1.4 Key Terms

To facilitate clarity through the remainder of the document, this subsection provides definitions of important concepts in the current research. These definitions capture the meanings most relevant to the context of the current work. The terms defined here are *health advice*, *RCTs*, and *observational studies*.

1.4.1 Health Advice

Prasad et al. (2013) defined health advice as recommendations related to any activity that might be performed by members of a health care team. They gave binary labels to research articles designating whether they provided health advice or not. Sumner et al. (2014) annotated health advice at the sentence level and further distinguished health advice as either “explicit” or “implicit” type (as shown in Table 1). By their definition, *explicit advice* is linguistically characterized by a direct recommendation for health-related behavior changes. In comparison, *implicit advice* hints at changes without making a direct recommendation, and thus may use different linguistic cues. Furthermore, *explicit advice* indicates a higher level of certainty than *implicit advice*, since straightforward recommendations are made for behavioral change. Read et al. (2016) annotated recommendations in clinical practice guidelines based on their strength. They categorized advice as “strong”, “moderate”, or “weak” to indicate its importance and the level of confidence of the advice giver (as shown in Table 2). To capture nuanced differences in language expression, we will apply these classifications used by Sumner et al. (2014) and Read et al. (2016). A detailed description of the categorization will be given in Chapter 4.

Table 1: Examples of health advice by its explicitness.

Explicitness	Examples
Implicit	1. “MMP-1 causes matrix destruction in TB, and therefore we believe it represents a novel therapeutic target to limit immunopathology.” 2. “Mid-late childhood (around age 7-11 years) may merit greater attention in future obesity prevention interventions.”
Explicit	3. “…[E]very patient needs to bring over-the-counter and prescription drugs to their doctor’s appointment for a comprehensive review.” 4. “We would advise people who want to drink sugar-sweetened beverages should do so only in moderation.”

Table 1: Examples of recommendations from clinical practice guidelines, rated by their strength.

Strength	Examples
Weak	1. “Obesity (body mass index [BMI] greater than 30kg/m2) is a condition for which there is no restriction on the use of the progestogen-only implant.”
Moderate	2. “Clinicians might offer Sativex oromucosal cannabinoid spray (nabiximols), where available, to reduce symptoms of spasticity, pain, or urinary frequency, although it is probably ineffective for improving objective spasticity measures or number of urinary incontinence episodes.”
Strong	3. “Assess for deterioration of the ulcer or possible infection when the individual reports increasing intensity of pain over time.” 4. “TEE should be performed in patients considered for percutaneous mitral balloon commissurotomy to assess the presence or absence of left atrial thrombus and to further evaluate the severity of mitral regurgitation (MR).”

1.4.2 RCTs

A *RCT* is a type of study design that randomly assigns individuals to experimental and control groups. The effects of treatments or interventions on the experimental group are compared to the effects on the control group (Kabisch et al. 2011). With the increasing importance of evidence-based medicine, RCTs are regarded as the best way to study new treatments and interventions in clinical research (Faraoni and Schaefer 2016).

1.4.3 Observational Studies

An observational study is a type of study in which individuals are observed or certain outcomes are measured. In observational studies, no interventions and treatments are carried out by researchers to affect the outcome (Mann, 2003). Observational studies are widely applied in the fields of epidemiology, social sciences, and psychology when RCTs are not always possible or cannot be conducted ethically (Song and Chung, 2010). Common types of observational studies include cross-sectional, case-control, retrospective, and prospective studies. *Cross-sectional studies*, also known as *prevalence studies*, analyze the number of cases in a population or a representative subset at one point in time (Mann, 2003). *Case-control studies* are designed to investigate risk factors that may prevent or cause the outcome. The design involves the comparison of participants affected by an outcome (cases) with a group of participants who are free of the outcome (controls) (Schlesselman, 1982). *Retrospective* and *prospective studies* are collectively referred to as *cohort studies*, which are used to measure events in chronological orders. Retrospective cohort studies investigate the past to examine events or outcomes observed in the past; in contrast, prospective cohort studies are performed with an eye toward the future and measure a variety of variables that might be related to the outcome (Song and Chung, 2010).

Among the four types of observational studies, cross-sectional studies are used to identify prevalence, while the other three types seek to identify risk factors and potential causal relationships (Mann, 2003). In general, cross-sectional studies are weaker than case-control, retrospective and prospective cohort studies in terms of study designs (Song and Chung, 2010; Murad et al., 2016).

1.5 Document Organization

With the background, research goal, and purpose of the study laid out in the current chapter, the remainder of the document is structured as follows: Chapter 2 presents a literature review of related work. Chapter 3 presents the research questions and methodology of this work. Chapter 4 addresses RQ1. Chapter 5 presents the analyses and results of RQs 2-5. Chapter 6 discusses the findings of the current work and directions for future work.

Chapter 2 Related Work

Detecting health advice in medical research papers consists of two subtasks: identifying statements that give advice and categorizing the advice by its level of commitment. As an inherent form of imperative language, advice conveys a speaker’s wishes or suggestions about an action (Condoravdi and Lauer, 2012). The level of commitment indicates the strength of the advice. It is normally manifested by language indicators, such as hedges, modalities, and evidentials. This chapter first reviews the linguistic foundations of, and computational approaches to, health advice detection. Second, it introduces the issue of problems with the quality of health advice in the medical literature, which applies the rationale for developing computational approaches to external quality assessment and evaluation.

2.1 Language Foundations and the Computational Modeling of Advice

2.1.1 Indicators Used in Expressing Advice

2.1.1.1 Linguistic Foundations of Advice

Advice is essentially a form of imperative language that functions as an illocutionary act. According to Austin (1962), illocutionary acts are observed when someone delivers a finding, gives or commits a decision in favor of or against an action, presents or explains views, or expresses reactions to other people’s behavior and attitudes. Searle (1976) further separated illocutionary acts into five categories: representatives, directives, commissives, expressives, and declarations. Among these categories, *directives* indicate the attempts that a speaker would like the hearer to do something. The approach can be very modest, for instance when an invitation or suggestion is offered, or it can be very fierce, such as when the speaker insists the hearer perform a certain act.

According to these definitions and categorizations, imperative language is a subgroup of directives (Ervin-Tripp, 1976). However, an imperative utterance can function more than just as a directive; it can have a wide range of uses. For instance, Condoravdi and Lauer (2012) classified the functions of imperative languages into four groups: directives, wish-type uses, permissions and invitations, and disinterested advice. *Directives* refers to the imperatives that are intended to get the hearer to do something or refrain from doing something. Common forms of directives are commands, warnings, advice, and pleas. *Wish-type uses* are imperatives that express a speaker's wish. *Permissions* and *invitations* express a speaker's desire; they are commonly seen in form of permissions/concessions, offers, and invitations. In contrast, *disinterested advice* is a special class of advice wherein the speaker has no interest in the fulfillment of the imperative. Unlike directives, disinterested advice tries to entice the hearer by implication to act on the content.

2.1.1.2 Computational Modeling of Advice

Several research areas and applications in the NLP field are related to the computational modeling of advice; these include imperative detection and suggestion mining.

Imperative detection focuses on developing computational techniques to identify imperative language. Datasets such as email conversations (Kwong and Yorke-Smith, 2009), Wikipedia discussions (Mao et al., 2014), and TV show dialogue (Xiao, Slation, and Xiao, 2020) were built to develop and evaluate automated approaches to imperative detection. Some language data resources also have labels and tags for imperative language. For example, English Web Treebank (Bies et al., 2012) contains both formal and informal texts extracted from weblogs, reviews, question-answer pairs, newsgroups, and emails. All the sentences were manually annotated for syntactic structures (e.g., POS tagging), and imperatives are included in the annotation.

Compared to the number of efforts made with other NLP applications, only a small number have been made to detect imperatives, and the majority of prior studies used rule-based approaches. For example, to detect imperatives in question-answer pairs extracted from email conversations, Kwong and Yorke-Smith (2009) applied naïve approaches using the regular expressions and algorithms of S&M (Shrestha and McKeown, 2004) in Ripper (Cohen, 1995). Mao et al. (2014) proposed two rules for extracting imperatives from Wikipedia's discussions. The first was to apply a dependency structure; specifically, if a verb was the root of a sentence and was in its base form with no subject child, the sentence was imperative. The second rule was that if the sentence had a modal verb with a personal pronoun or a noun as the subject, it was also imperative (as shown in Figure 1). Gupta et al. (2018) also applied a rule-based approach to extract imperative language. They used a pre-trained, rule-based parser featuring domain-specific words to detect the imperatives in technical documents.

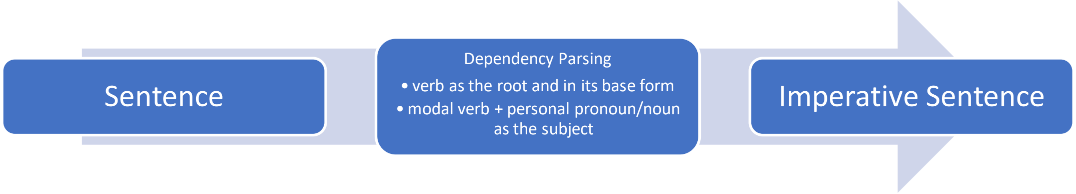

Sentence
Dependency Parsing
• verb as the root and in its base form
• modal verb + personal pronoun/noun as the subject
Imperative Sentence

Figure 1: An illustration of a rule-based approach for detecting imperative sentences.

Besides the task of imperative detection, suggestion mining, a research area in the NLP field, is relevant. Prior studies defined suggestion mining as a sentence-level classification task whose purpose is to detect wishes, advice, and recommendations in opinionated text (e.g., Goldberg et al., 2009; Ramanand, Bhavsar, and Pedanekar, 2010; Brun and Hagege, 2013; Negi, 2016; Negi, Daudert, and Buitelaar, 2019). To this end, different types of opinionated text such as customer reviews (e.g., Goldberg et al., 2009; Ramanand, Bhavsar, and Pedanekar, 2010; Brun and Hagege 2013; Negi 2016; Negi, Daudert and Buitelaar 2019), discussion forum posts (Goldberg et al., 2009; Wicaksono and Myaeng 2012, 2013), and tweets (Dong et al. 2013) were built.

Meanwhile, some corpora have been built for the extraction of sentences with functions like advice. For example, Read et al. (2016) developed a corpus of clinical guidelines annotated with annotated instances of recommendations. The guidelines were obtained from the National Guideline Clearinghouse, which is a public database maintained by the Agency for Healthcare Research and Quality. The strength of the importance of each recommendation is also indicated, as specified in the guidelines.

To automatically extract suggestions, both rule-based and machine-learning approaches have been used. Earlier work with suggestion mining adopted a rule-based approach to identify sentences with suggestions. This type of study often employed domain-specific and hand-crafted linguistic rules to extract advice-related statements (e.g., Ramanand, Bhavsar, and Pedanekar, 2010; Brun and Hagege 2013). In addition, machine-learning approaches, such as Conditional Random Fields (CRF) (Wicaksono and Myaeng 2013), Factorization Machines (Dong et al., 2013), and Support Vector Machines (SVM) (Negi and Buitelaar, 2015), have been utilized to identify suggestions, and their performance has been compared.

Recently, deep-learning approaches have also been used to identity sentences with suggestions. For example, in the suggestion-mining task of SemEval-2019 (Negi, Daudert and Buitelaar, 2019), models based on Convolutional Neural Networks (CNN) (e.g., Park et al., 2019; Yue, Wang, and Zhang, 2019) and Long Short-Term Memory (LSTM) (Cabanski, 2019) were developed to extract suggestions from online reviews and forums. Pre-trained language models, such as Bidirectional Encoder Representations from Transformers (BERT) (Devlin et al., 2018), have also been used to detect suggestions (e.g., Liu, Wang, and Sun, 2019; Park et al., 2019).

To date, the datasets that are available consist mostly of online customer reviews, social media posts, and clinical guidelines. Corpora on advice in scientific literature are lacking, especially in the health domain. Because rule-based and machine-learning approaches use different datasets, the experimental results of automated approaches might not be comparable, and their generalizability remains an open question for detecting health advice in medical research literature. Overall, suggestion mining remains an emerging research area in comparison to other NLP tasks. Health advice has not been computationally modeled as a language construct. Therefore, more work is needed to examine the feasibility of applying NLP techniques to the detection of health advice in scientific communication.

2.1.2 Indicators for Expressing Level of Commitment

2.1.2.1 Linguistic Foundations of Level of Commitment

Level of commitment shows how strong a statement is. To indicate the commitment, language indicators such as hedges, modalities, and evidentials are commonly used. This subsection reviews the linguistic foundations of these indicators.

Lakoff (1972) defined hedges as “words whose meaning implicitly involves fuzziness – words whose job is to make things fuzzier or less fuzzy” (p.195). Myers (1989) argued that a claim that has no hedging is probably not a statement of new knowledge. In fact, despite a widely held belief that professional scientific writing should consist of impersonal statements of fact which add up to the truth, hedges have been found to be abundant in scientific discourse. Hedges can express tentativeness and possibility, indicate the level of commitment writers attach to their statements, and qualify an author’s confidence in the truth of a proposition (Hyland, 1998a; 1998b).

The purpose of hedging is not only to modulate the epistemic validity that writers have inferred from given evidence; hedges can also be used to signal vagueness, evasion, equivocation, and politeness (Fraser, 2010). Specifically, *vagueness* refers to a situation where received information lacks the expected precision. *Evasion* occurs when the information fails to meet expectations. *Equivocation* is the use of words with more than one meaning with the intention of misleading the hearer; it is a type of non-straightforward communication that is ambiguous, contradictory, or even evasive (Bavelas et al., 1990). Mauranen (2004) referred to hedging as a pragmatic phenomenon and connected it to politeness. The boundary of a hedge is then extended to "negative politeness" (Brown and Levinson, 1978), which is used to avoid threats to the dignity of the participants. For instance, hedges are used as a rhetorical strategy of politeness in science communication to minimize the potential threat of a new claim to peer researchers (Myers, 1989). Recognizing reviewers as disciplinary gatekeepers, writers may observe community expectations of self-assurance but use hedges to "negotiate" the claims (Hyland, 1998a; 1998b). Therefore, hedging may be used to display not only a writer's or speaker's degree of confidence, but also to indicate how much confidence they feel it is appropriate to display (Crompton, 1997).

Given the critical role of hedging, many studies have attempted to define its scope and create taxonomies of the linguistic devices used for hedging. To date, a wide range of lexical, grammatical, and strategic devices have been considered as hedges (Hyland, 1998a; 1998b). For instances, Zuck and Zuck (1986) proposed a list of linguistic devices that are usually used for hedging, which includes auxiliaries (e.g., *may*, *might*, *could*), semi-auxiliaries (e.g., *seem*, *appear*), full verbs (e.g., *suggest*), passive voice, adverbs, and adverbials (e.g., *probably*, *relatively*, *almost*), adjectives, indefinite nouns, and pronouns. Markannen and Schroder (1987) provided a similar list but added a few specifications of their own. They claimed that, apart from the list, the use of one word and avoidance of another, as well as the choice of a specific vocabulary, could also be treated as an instance of hedging.

Salager-Meyer (1994) analyzed a corpus of 15 articles in leading medical journals and proposed a taxonomy of five categories of hedges (as shown in Table 3). Namasaraev (1997) identified four parameters that characterize the hedging strategies (as shown in Table 4). Heng and Tan (2000) also proposed a taxonomy, but unlike the two schemas just mentioned, theirs was similar to the one proposed by Zuck and Zuck (1986), except that they) included the descriptions of where hedges normally occurred.

Table 2: Salager-Meyer's taxonomy of hedges (1994).

Type	Definition	Examples
Shields	all model verbs expressing possibilities, semi-auxiliaries, probability adverbs and their derivative adjectives, epistemic verbs	appear, seem, probably, suggest
Approximators	adaptors and rounders of quantity, degree, frequency and time	approximately, roughly, somewhat
Author's personal doubt and direct involvement	phrases showing personal doubt and involvement	I believe, to our knowledge
Emotionally charged intensifiers	phrases indicating strength or intensity	extremely difficult, particularly encouraging
Compound hedges	strings of hedges	it may suggest that

Table 3: Namasaraev's categorization of hedge strategies (1997).

Type	Definition	Examples
Indetermination	adding a degree of uncertainty or fuzziness to an utterance	appear, seem, probably
Depersonalization	avoiding direct reference	we, researchers, authors
Subjectivization	using a personal pronoun + a verb of thinking; signaling the subjectivity of a term; noting that a statement is only an opinion, rather than the absolute truth	I + think/suppose/assume
Limitation	eliminating vagueness or fuzziness with a limitation	

Although the abovementioned taxonomies share a few similarities among the commonly used hedging strategies, unanimous agreement is lacking regarding the forms and functions of hedges. According to a summary by Crompton (1997), the categories of hedging devices that are recognized by multiple researchers are lexical verbs (e.g., *suggest*), modal verbs (e.g., *might*), probability adverbs (e.g., *perhaps*), and probability adjectives (e.g., *possible*). Some categories are less agreed upon, such as if-clauses and approximators (e.g., *roughly*).

Another group of linguistic devices that is linked to hedges is modalities. A *modality* is an expression of an individual's subjective attitude or opinions (Bybee and Fleischman, 1995). It is a semantic domain of elements of meaning expressed by language, and it covers a broad range of semantic nuances (e.g., jussive, desiderative, hypothetical, potential, and dubitative) (Bybee and Fleischman, 1995). The linguistic understanding of modality was derived from modal logic. The term traditionally associated with level of commitment and speculation in language is *epistemic modality* (Palmer, 1986).

Epistemic modality is reflected in *epistemic comments*, described as a writer's assessments of the possibilities expressed in a statement. Most linguists considered epistemic modality to be an indication of a speaker's judgment of the truth in a proposition and the speaker's attitudes toward it. For instance, Coates (1987) stated that epistemic modality reflects a speaker's level of confidence in the truth of a proposition. Halliday (1970) described epistemic modality as a speaker's assessments of probability and predictability. According to his description, the speaker's assessments, carried by epistemic modality, is external to the content but shows the speaker's attitude toward his own speech. Similarly, Palmer (1986) argued that epistemic modality indicates "the status of the proposition in terms of the speaker's commitment to it" (p. 54-55). Bybee and Fleishman (1995) noted that epistemics are clausal-scope indicators of a speaker's commitment to the truth of a proposition.

Other than functioning to indicate strength of an expressed proposition, epistemic modality can also be pragmatically applied as a politeness strategy, a face-saving strategy, or a persuasion and a manipulation strategy (Kärkkäinen, 1992). Through epistemic modality, speakers can establish a relationship with the addressees by presenting their ideas and thoughts in a more polite manner (Yang et al., 2015). In addition, cultural background can influence speakers' use of epistemic modalities for different functions (e.g., Youmans, 2001).

Traditionally, the study of epistemic modality has been confined to modal auxiliaries (e.g., Palmer 1986; Kärkkäinen, 1992), but more recently a wider view has been adopted, which includes other parts of speech (e.g., Rizomilioti, 2006). Epistemic modalities can range from expressions of uncertainty to certainty, through various language features like adjectives (e.g., *probable*, *potential*, *possible*, *certain*, *definite*, *clear*), adverbs (e.g., *impossibly*, *positively*, *possibly*, *scarcely*, *certainly*), verbs (e.g., *can*, *could*, *have to*, *must*, *might*, *should*), and nouns (e.g., *chance*, *opportunity*, *possibility*).

The critical role of epistemic modality in writing, especially in scientific discourse, has been examined by previous researchers (e.g., Hyland, 1994, 1995, 1996; Hu and Cao, 2011; Wharton, 2012), who have mostly analyzed the frequency of certain modal words in texts, as well as their functions. Studies have suggested that in academic discourse, epistemic modalities frequently occur in introduction and discussion sections and are less frequent in the results and methods sections (Hyland, 1994, 1995). Also, epistemic modalities have been found to be especially frequent in the conclusion, recommendation, and data synthesis sections (Salager-Meyer's, 1992).

Level of commitment can also be reflected in the use of evidentials, which indicate a degree of information reliability. By Anderson's (1986) narrow definition, an *evidential* states the evidence a person has for making a factual claim. It normally refers to linguistic devices used for subjective relations and knowledge. Based on this definition, linguistic studies of evidentiality are primarily concerned with the evidential forms and meanings in morphological systems and focus on languages other than English (Mushin, 2001). By a broader definition (Chafe and Nichols, 1986), evidentials involve various attitudes toward knowledge and their functions extend beyond marking the evidence in for a claim. They can evaluate the degree of reliability of knowledge, specify the mode of knowledge, and mark a contrast between knowledge and expectation. In this sense, evidentials are concerned with expressions of truth, doubt, reliability, confidence, and authority, and many other elaborations of people's attitudes (Mushin, 2001) and they are closely related to the expressed level of commitment in language.

Although English does not have a specific category of evidentials, previous studies have reached the conclusion that English compensates for this lack by other means. A variety of words and expressions can function as evidentials, including modal words, adverbs, conjunctions, prepositional phrases and predictions (e.g., *I believe that, he claims that*) (Barton, 1993).

Based on the current understanding, evidentials and notions like hedges and epistemic modalities share many similar words and expressions. Many epistemic verbs, and hypothetical constructions that are commonly used in hedging and epistemic comments are also frequently used as evidentials. However, people with different language backgrounds and competencies may have different styles of expressing evidentials meanings and functions. For example, Barton (1993) conducted a discourse analysis on the use of evidentials in 100 essays written by experienced academic writers and 100 essays written by student writers from a variety of academic fields. The comparison showed that experienced academic authors took advantages of evidentials to specify their purposes, theses and arguments. Neff et al. (2003) compared the use of evidentials between native and nonnative English writers and noted that the differences were significant. Nonnative speakers overused *can* in comparison to native speakers, and they underused modal words such as *might*, *may*, and *could* in the writing.

2.1.2.2 Computational Modeling of Level of Commitment

The NLP tasks of detecting speculative statements and hedges share many similarities with modeling level of commitment. Previous work on speculative statement detection mostly focused on classifying sentences into speculative or definite categories and on detecting the scope of speculative statements. Current approaches are mainly based on supervised methods using different feature engineering methods.

Light et al. (2004) developed a classifier to predict the speculative sentences in biomedical abstracts. They built a rule-based system by using 14 predefined strings (i.e., *suggest*, *potential*, *likely, may, at least*, *in part, possible, potential*, *further investigation*, *unlikely*, *putative*, *insights*, *point toward*, *promise*, *propose*), which outperformed the one using SVM with stemming and term-frequency representation, and a baseline model using majority vote. The better performance over the use of SVM and substrings suggests that the speculative language in their sampled abstracts could be detected through the shallow lexical features.

Wei et al. (2013) also used SVM to detect the uncertainty in tweets. In addition to the n-gram features, they added content-based, user-based, and twitter-specific features. *Content-based features* included length (the length of tweets), cue phrases (whether the tweets contained certainty cues or not), and the ratio of words out of vocabulary. *Twitter-specific features* included the existence of a URL, the frequency of URLs in the corpus, the times of retweets, the occurrence of hashtags, the number of hashtags on the tweets, and information on whether the current tweet was a replay or retweet. *User-based features* mostly described the numbers of followers, lists, friends, favorites, and tweets a user had. Among all the feature representations, SVM with n-gram representations and all the three types of additional features had the highest F1-score.

Yang et al. (2012) used CRF with a wide range of linguistic features to recognize speculative sentences in requirement documents. The linguistic features included word-token features (e.g., part-of-speech tagging, the chunk-tagging of word), context features (e.g., trigram features), dependency relation features, and co-occurrence features. In addition, they also considered many linguistic cues related to uncertainty expressions, such as auxiliaries, epistemic verbs, epistemic words, and conjunctions, to detect the scope of speculative sentences.

Besides the classic algorithms such as SVM and CRF, some other algorithms and approaches have also been applied. For example, Szarvas (2008) developed a Maximum Entropy classifier that incorporates bigrams and trigrams into the feature representation. Li et al. (2014) formulated the task of speculative language detection as a sequence-labeling problem to capture the dependency between neighboring words, and they applied the classical Hidden Markov Model (HMM) with a specific tag set to label a sentence at the word level. They applied the model to the BioScope corpus (Vincze et al., 2008) and to the Wikipedia dataset used in CoNLL-2010. CNN and Recurrent Neural Networks (RNN) have also been applied to speculation detection. For example, Adel and Schütze (2017) applied CNN and RNN to uncertainty detection and compared their performance with several baseline models, using the same dataset as Li et al. (2014) did. They found that both CNN and RNN outperformed the model built on SVM.

Hedge detection is also a related area. Light et al. (2004) conducted one of the earliest studies of automated hedge detection. With an annotated corpus of hedging cues in biomedical documents, they performed the first experiment in automatic hedging and speculation classification. Medlock and Briscoe (2007) modeled hedge classification as a weakly supervised machine-learning task, on articles from the functional genomics literature. They developed a probabilistic classifier to acquire training data, starting with a small set of seed examples to indicate hedging and then iterating more training seeds without much manual intervention. Medlock (2008) later extended the work by using more features, such as part-of-speech tagging, stemming, and bigrams. The experimental results suggest that stemming improves the performance of the model and that the best results are obtained with stemmed unigram and bigram representations. Following the above exploration, Szarvas (2008) developed a Maximum Entropy classifier that added trigrams to the feature representation to perform a reranking-based feature selection procedure, which reduced the number of keyword candidates. By training their system using the same dataset as Medlock and Briscoe (2007) and testing on newly annotated biomedical articles and clinical reports, Kilicoglu and Bergler (2008) applied a linguistically motivated approach to the same classification task, using syntactic patterns and knowledge from lexical resources. In their experiment, hedge cues were weighted by information gain measures and by weights assigned according to their types and centrality to hedging.

Morante and Daelemans (2009) developed a two-phased approach to detect the scope of a particular hedging cues in biomedical articles. They divided the detection of certain linguistic cues and their scope in the text into two separate steps. The F1-scores for identifying hedging scope in abstracts, full-texts, and clinical articles were higher than those for the baseline approach of tagging dictionary words as hedging cues.

Agarwal and Yu (2010) trained a model using CRF on the BioScope corpus. They marked each word in the corpus to indicate whether it was part of the hedge cue or not. Specifically, the first word in a hedge cue was marked to indicate the beginning of the cue, and the remainder of the hedge phrase was marked to indicate the body of the cue. The trained model was then used to automatically identify hedge cues in the test sentences by tagging the first word and those that followed. The scope of the hedge cues was marked in a similar way. For the scope detection, they incorporated part-of-speech tagging to resolve the clause issues that might confuse the hedge scope identification. The best CRF model performed significantly better than the baseline system of marking the hedge cues and punctuations.

Hedge detection was also one of the CoNLL-2010 shared tasks (Farkas et al., 2010). The shared tasks included two phases: (1) detecting the propositions containing uncertainty at the sentence level and identifying the hedge cues and (2) detecting the linguistic scope of hedge cues in sentence. The best system for Wikipedia data employed SVM (Georgescul, 2010) and the best system for biological data adopted CRF (Tang et al., 2010). Tang et al. (2010) trained both a CRF sequence classifier and an SVM-based HMM model, finally combining the predictions of both models in a second CRF to make predictions. Among all the submissions, the approaches applied included sequence labeling, token classification, and bag-of-words models and several machine-learning approaches were used, such as Entropy Guided Transformation Learning, Averaged Perception, the k-nearest neighbors algorithm, CRF, HMM, and SVM. Features representations like lemmatization, stemming, part-of-speech tagging, and dependency relation were also used to train the model.

Following the CoNLL-2010 shared task, Velldal (2011) proposed modeling the hedge detection task as a disambiguation problem, focusing on words that had previously been identified as hedge cues; this greatly reduced the number of examples for the feature space. Velldal (2011) built a large-margin SVM classifier with n-gram features in addition to the part-of-speech tagging, lemmatizations, and other shallow representations; the model built outperformed the one developed by Tang et al. (2010).

From these studies, it can be seen that biomedical data have been commonly used to train the models for hedge detection. Even if the detailed approaches and algorithms applied for the different tasks were not identical, nearly all of them were linguistically motivated, using patterns specified by hand-crafted rules or other supervised learning approaches.

2.2 Problems with the Quality of Health Advice in the Medical Research Literature

Research publications are an integral part of the scientific process. They introduce new knowledge and concepts and communicate scientific information among scientists and with the general public. Therefore, accurately representing procedures and findings plays a critical role in science communication (Kleinert and Wager, 2010). However, scientists are concerned about the quality of information in the medical research literature. In particular, medical experts have warned about the problem of misinterpreting research results and their implications, which can mislead readers to view a study in a more favorable light than is warranted (Boutron et al., 2010; Ochodo et al., 2013; Lazarus et al., 2015; Chiu, Grundy, and Bero, 2017; McGrath et al., 2017; Boutron and Ravaud, 2018).

Among all kinds of misinterpretation, inaccurately making health advice from observational studies in medical literature is identified as one common type. For example, Prasad et al. (2013) found that health advice inferred from observational study findings is common in medical publications. Based on their manual examination of about 300 observational studies in leading medical journals, they noted that about 56% contained advice. Wilson and Chestnutt (2016), through a content analysis of peer-reviewed dental journals, found about 30% of the advice relating to clinical practice was not fully supported by the study presented. Although there are arguments both for and against the value of observational studies for informing and guiding health-related decisions, the advice found there frequently involves logical leaps that lead to possible misinterpretations of study findings (Prasad et al., 2013; Banerjee and Prasad, 2020).

Establishing advice based on research evidence is not a trivial task (Brown et al., 2006; Shah et al., 2017). The complexity of research designs can be a major challenge for scientists with inadequate training to give proper advice that can be justified by study designs (Thiese, 2014). The EBM Pyramid (Glover et al., 2006) is often used to delineate the quality of research evidence from various types of study designs. Overall, the evidence is grouped into two main types: unfiltered information from primary studies and filtered information from secondary studies, such as systematic reviews and practice guidelines. The primary studies are of multiple design types with different strength levels (Thiese, 2014), including RCTs and various types of observational studies. The quality of evidence can reflect the degree of certainty or confidence in the estimates of effects in relation to an outcome, which then influences the trustworthiness and strength of the recommendations that can be made (Woolf et al., 2012).

At the same time, scientists may also have a tendency to rely on their "wishful thinking" rather than scientific evidence to draw research conclusions, as psychology studies have found that people, including scientists, can incline to the beliefs that they want to accept (Coyne and Tennen, 2010). Theories from the disciplines of decision science, health psychology, and communication generally agree that people rely on two inter-related systems during the reasoning process – an experiential-automatic process (system 1) and an analytic-deliberative process (system 2; see, e.g., Green and Brock, 2000; Butow et al., 2007; Volk et al., 2008). System 1 is quick and effortless, while system 2 includes more active reasoning, which is effortful and cognitively demanding (Bekker et al., 2013). The process of deriving recommendations from research evidence should be analytic and deliberative, requiring cognition and scientific reasoning. However, scientists can be "motivated reasoners" highly influenced by their hopes and emotions when evaluating evidence (Halpern, 1998). As result, if the evidence is congruent with their prior beliefs, there is a tendency to rely on the experiential-automatic process, rather than the analytic-deliberative one, to draw conclusions from a study.

Another factor that may further challenge the practice of giving advice is the wording used for recommendations. Comprehensive evaluations of recommendations in clinical practice guidelines suggest a lack of standards for wording recommendations, and recommendations were presented with great inconsistency among different clinical guidelines (Schünemann, Fretheim, and Oxman, 2006; Woolf et al., 2012). Although quite a few instructions and frameworks, such as

GRADE (Grading of Recommendations Assessment, Development, and Evaluation) (Andrews et al., 2013), PICO (Population, Intervention, Comparison, and Outcomes), and EPICOT (Evidence, Population, Intervention, Comparison, Outcome, and Time) (Brown et al., 2006) have been proposed to help scientists formulate evidence-based guidelines and recommendations for clinical practice, clinical recommendations have commonly been found to have problems with clarity (McDonald and Overhage, 1994; Shekelle et al., 2000; Michie and Johnston, 2004).

The challenge of making health recommendations is further exacerbated by competition for scientific impact and reputation, which may lead scientists to overinterpret their study results and inappropriately enhance the implications of study findings (Ioannidis, 2005; Chiu, Grundy and Bero, 2017; Boutron and Ravaud, 2018). This overinterpretation can result in the excessive promotion of scientific developments and applications. For example, empirical studies have identified three common types of overinterpretation in science communication: the misrepresentation of causal claims associated with correlational study findings (Robinson et al., 2007; Cofield et al., 2010; Yu et al., 2019), claims about effects on humans inferred from animal studies (Sumner et al., 2014; Chang, 2015), and the extrapolation of exaggerated health advice (Prasad et al., 2013; Haneef et al., 2015; Lazarus et al., 2015). Overstated research claims often occur not only in news stories (Haneef et al., 2015; Sumner et al., 2014), but also in the research literature, particularly in the health and biomedical domains (e.g., Cofield et al., 2010; Prasad et al., 2013; Lazarus et al., 2015).

Meta-analyses have also found that overstatements are more likely to occur in abstracts than in the body text of medical research publications, and thus, it is suggested that clinicians and policy makers not rely solely on the advice in abstracts for decision-making (Assem et al., 2017; Nasciment et al., 2021). Furthermore, the location of advice can affect its reach and impact. Advice given in abstracts was found to be farther reaching – if an abstract discussed the significance of a studyand hinted at a change in practice, clinicians were more likely to read the full text and rate a treatment as beneficial (Boutron et al., 2014). The fact that advice can be given in more than one place also can lead to problems with consistency. For example, a content analysis of systematic reviews of therapeutic interventions found that a large proportion of health advice in abstracts was not only overclaimed but also inconsistent with the advice given in the discussion sections, where researchers would have more room to explain their advice in more detail (Yavchitz et al., 2016).

2.3 Summary

This chapter reviews the linguistic foundations of imperative language and level of commitment. It also reviews the computational approaches to modeling health advice-related language phenomena, including the detection of imperatives, suggestions, speculative statements, hedges, negations, and contradictions. The chapter also discusses problems with the quality of health advice in medical research publications. It focuses on instances of exaggeration in research papers, when health advice cannot be fully supported by the research results and findings.

Overall, linguistic foundations provide us with good insights for use in health advice analysis. However, the majority of current work is based on small sample sizes, and norms and rules derived from observations may be applicable only to certain situations. Although promising levels of accuracy have been achieved, several aspects of current computational approaches still need further investigation. Training corpora and computational models particular for medical research literature are needed for detecting health advice in medical research literature.

Chapter 3 Research Questions and Methodology

This thesis aims to advance the automated identification of health advice in scientific literature. The study was expected to yield a deeper understanding of health advice as a language construct, and to broaden our knowledge of advice giving in medical research articles, especially that found in observational studies. The overall design of the study consists of two parts, with the development of advice taxonomy, annotation corpus and prediction model as precursor to the examination of health advice offered in medical research papers. The taxonomy, corpus, and successful prediction model developed in the first part will lay the foundation to answering research questions regarding the practice and prevalence of giving health advice in the medical research literature.

In the first part, the study addresses *RQ1*: To what extent can NLP prediction algorithms detect health advice in PubMed publications? Since health advice has not been computationally modeled as a language construct before, we lack available datasets and NLP techniques for advice detection in medical research papers. To answer this question, the study first developed an annotated taxonomy and corpus of health advice to serve as the gold standard dataset. The study designed and evaluated both traditional machine-learning and deep-learning approaches for classifying health advice on the corpus.

The annotated training corpus and a successful prediction model developed in the first step will lay the foundation for answering the following research questions:

RQ2: Where do research papers give health advice? If a research paper gives health advice in both the abstract and the discussion, are the advice statements equivalent?

RQ3: Is health advice prevalent in observational studies? How have patterns changed over time?

RQ4: Do journals differ in their practice of allowing advice giving or not?

RQ5: What health advice has been offered regarding the use of HCQ for treating COVID-19?

RQ2 examines the practice of giving advice in abstracts and discussions. Although prior studies have raised concerns over the advice-giving behavior of medical researchers, in-depth investigations are lacking, leaving many questions unanswered. For example, what is the major practice in individual studies, to give advice or not? Where do researchers often give advice, in the abstract or in the discussion? If recommendations are given in both sections, are they semantically equivalent or does one version tend to be stronger than the other? Utilizing the annotated corpus from RQ1, a content analysis was conducted to determine the number of papers that gave advice versus those that did not and to assess the strength of the advice given in abstracts versus discussion sections.

RQ3 concerns the prevalence of health advice in observational studies. Prior studies, such as one by Prasad et al (2013), argued that the prevalence of health advice in research papers, especially observational studies, raises concerns about scientific rigor. However, prior estimations were based on manual content analyses on small samples of articles. Due to the large number of research outputs, such a labor-intensive approach is difficult to maintain. Manual content analysis is also inadequate for answering important research questions that require large-scale analyses. For example, how prevalent is health advice in observational studies over the years? Utilizing the NLP technique for health advice detection developed in response to RQ1, this stage (RQ3) reexamines the prevalence of, and trends in, giving health advice in observational studies on a large scale over time.

RQ4 focuses on the relationship between journal impact and the prevalence of advice in observational studies. As described in Chapter 1, prior content analyses on research papers of specific health topics show opposite results regarding the relationship between journal impact and the prevalence of health recommendations or interpretations of clinical applicability (e.g., Lumbreras et al., 2009; Wilson and Chestnutt, 2016). Whether journals differ in their practice of allowing advice giving in observational studies is still unknown. To answer this question, this study applied the developed advice prediction model to observational studies in PubMed and adopted the generalized linear model (McCullagh and Nelder, 1983) to examine the relationship between advice giving and differences in journal impact in a large-scale dataset of medical research papers over the past 25 years.

RQ5 deals with health advice that has been given in the medical research literature regarding the use of HCQ to treat COVID-19. HCQ was considered a promising treatment option at the beginning of the COVID-19 pandemic but in later clinical trials was found to be ineffective. RQ5 asks whether the NLP technique developed for this study could be used to retrieve health advice on a specific medical topic, especially when used in combination with current health information services. An application case study was carried out, applying the prediction model for retrieving health advice to the case of HCQ as a treatment option, using LitCovid, a large COVID-19 research literature database curated by NIH.

3.1 Research Design: Part 1 – NLP Modeling of Health Advice

3.1.1 Developing a Health Advice Taxonomy

Drawing on health advice definitions from past studies (e.g., Prasad et al., 2013; Summer et al., 2014; Read et al., 2016), this study constructed an advice taxonomy that categorizes sentences in medical research abstracts and conclusions based on two aspects of advice: its occurrence and its level of commitment.

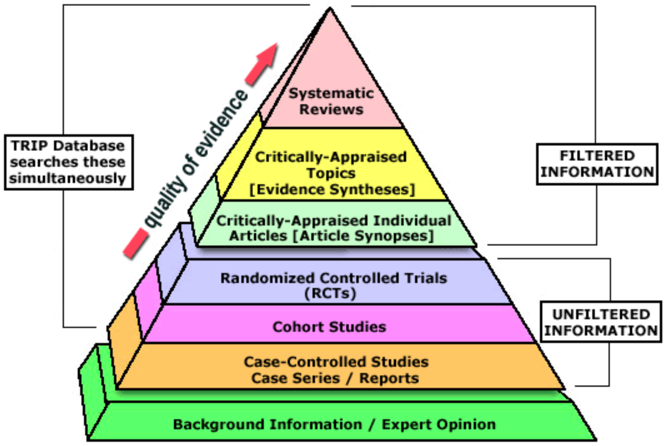
Occurrence indicates whether or not a sentence contains advice for a health-related behavior change. For this dimension, a sentence is annotated as either “advice” or “no advice”. A “no advice” statement describes study background, results, findings, limitations, or suggestions for future studies, and so forth, with no suggestion for a change in health-related behavior (e.g., behavioral, clinical or medical) pr a change in clinical practice. In comparison, an “advice” statement should suggest a change in health-related behavior. Advice statements also include clinical recommendations and policy-oriented call-for-action recommendations. More detailed definitions and examples are presented in Chapter 4.

Level of commitment refers to the strength of the advice. Based on Sumner et al.’s (2014) past analysis of the explicitness of health advice and Read et al.’s (2016) analysis of the strength of clinical guideline recommendations, level of commitment was categorized into two classes, “weak” and “strong”. A statement with weak advice hints that either a behavior or a health-related practice needs changing, or it suggests that there are certain options and alternative approaches to a current clinical or medical practice. A sentence with strong advice makes a straightforward advice recommendation regarding a health-related behavior or practice.

This study used the common categorical agreement measure Cohen’s kappa (Cohen, 1960) for inter-coder agreement testing. More detailed information on the annotation schema and inter-coder agreement checking will be presented in Chapter 4.

3.1.2 Constructing an Annotated Corpus of Health Advice

Constructing a reliable, hand-coded dataset is needed to serve as ground truth for testing the automatic recognition of health advice. In the current study, sentences were extracted from the medical research papers and annotated based on the annotation schema. PubMed[1] was selected as the data source. According to the EBM Pyramid (as shown in Figure 2), different study designs lead to different levels of evidence for medical decision making (Murad et al., 2016). To ensure that the health advice prediction model was effective for identifying health advice across study designs, a sample of 6,000 sentences was selected from both RCTs and observational studies by using MeSH terms in PubMed (Corpus-Train). The sample included four common subtypes of observational studies: cross-sectional, case-control, retrospective, and prospective studies, listed here in increasing order of evidence strength. The Stanford CoreNLP tool was used to preprocess and parse the downloaded XML files downloaded from PubMed.


TRIP Database searches these simultaneously
quality of evidence
Systematic Reviews
Critically-Appraised Topics [Evidence Syntheses]
Critically-Appraised Individual Articles [Article Synopses]
Randomized Controlled Trials (RCTs)
Cohort Studies
Case-Controlled Studies Case Series / Reports
Background Information / Expert Opinion
FILTERED INFORMATION
UNFILTERED INFORMATION

EBM Pyramid and EBM Page Generator, Copyright 2006-2014 Trustees of Dartmouth College and Yale University. All Rights Reserved. Produced by Jan Glover, David Izzo, Karen Odato and Lei Wang.

Figure 2: EBM Pyramid (Glover et al., 2006).

Three annotators, with backgrounds in information science, linguistics, and clinical psychology, annotated the sentences for types of health advice. During the annotation process,

[1] PubMed is the largest health literature database. Besides abstracts, it provides rich metadata that can distinguish research papers with different types of study designs.

language indicators of different types of health advice were highlighted. All ambiguous cases were brought to the team members for discussion.

3.1.3 Developing and Evaluating NLP Techniques for Health Advice Detection

Like tasks in suggestion mining, the current work frames the detection of health advice as a sentence-level text classification task. For the traditional machine-learning approach, the study measured the performance of SVM with different vectorization methods and enriched features to train the sentence-type classifiers, by using the Scikit-learn Python package and combining the SVM (Liblinear) algorithm with three different frequency measures – word presence and absence (SVM-boolean), word frequency (SVM-tf), and word frequency weighted by inverse document frequency (SVM-tfidf).

Recent developments in deep-learning techniques provides new methods such as BERT (Devlin et al., 2018), which can effectively learn local context and long sequences. BERT is a transformer-based machine-learning technique for NLP. Unlike other approaches, it processes the language input bidirectionally (from-left-to-right and from-right-to-left) at the same time. It enables parallelization and improves the performance of attention mechanism by introducing self-attention, which can understand context-heavy texts (Devlin et al., 2019; Fan et al., 2020). Therefore, this pre-trained language model has already learned many general linguistic patterns that can be further used in various NLP tasks by retraining the models with new training data for specific tasks. Such end-to-end approaches can save a huge amount of human effort in looking for specific linguistic patterns.

BERT has achieved state-of-the-art results on several NLP tasks. In a task such as suggestion mining, which is like the task here, the BERT-based transformer approach outperformed the other machine-learning approaches developed for the SemEval-2019 task (Negi et al., 2019).

Compared to BERT, BioBERT is further pre-trained on a large-scale biomedical dataset. It outperforms the original BERT model on biomedical named entity recognition, biomedical relation extraction, and biomedical question answering (Lee et al., 2020).

This study utilized the available existing BERT and BioBERT models that were trained on large-scale general-purpose corpora and improved them with the annotated data of health advice. The specific parameters used in the study include three epochs, a learning rate of 2e-5, and a max sequence length of 128 in the cased BERT-base model. The same BERT parameter settings were used for BioBERT, except with the utilization of the BioBERT pre-trained model rather than the cased BERT-based one.

All the prediction models were evaluated on the annotated dataset (Corpus-Train) with five-fold cross validation as the evaluation method. To evaluate the performance, macro-averaged precision, recall and F1 scores were reported. Error analyses were conducted to explain the patterns the models had failed to learn. As the training set was built on sentences extracted from structured abstracts, the generalizability of the model was also tested on a sample of data that contained both sentences from both unstructured abstracts and discussion sections in full text content (Corpus-Eval) in Chapter 4. The detailed evaluation process and results will be described in Chapter 4.

3.2 Research Design: Part 2 – Examining Health Advice Giving Behavior in Medical Literature

3.2.1 Comparing Advice Giving in Abstracts and Discussions

A content analysis of 100 research papers (Corpus-Eval) with abstracts and full-text content was conducted to compare the advice given in the abstracts and discussion sections. The study aggregated the sentence-level advice labels to sections levels to compare advice occurrence and level of commitment between the abstract and the discussion. The results determined whether the advice given in the abstract and discussion section of a paper was equivalent or not.

3.2.2 Measuring the Prevalence of Health Advice and Trends in Observational Studies

This study applied the prediction models to the observational studies (Corpus-Application) to examine the prevalence of, and trends in health advice, based on the ratio of health advice in research papers. This analysis focused only on abstracts, given that abstracts in PubMed are open access. Moreover, abstracts have been identified as the parts of medical research papers mostly affected by exaggeration (as described in Chapter 2). The National Library of Medicine produces an annual baseline dataset and an updated dataset of MEDLINE/PubMed citation records in XML format. Both datasets were downloaded from its FTP server on September 30, 2019, and all RCTs and observational studies were retrieved by using the MeSH terms (i.e. "Randomized Controlled Trials", "Cross-Sectional Studies", "Case-Control Studies", "Retrospective Studies", and "Prospective Studies") in the XML files. Articles with mixed-study designs were excluded.

The study applied the best-performing model to predict health advice in each sentence extracted from Corpus-Application. To evaluate the model's generalizability to all observational studies, 100 sentences from in the prediction result were randomly sampled and manually examined for accuracy. Sentence-level predictions were then aggregated to article-level predictions for and analysis of prevalence and trends over 25 years.

3.2.3 Examining the Relationship between Journal Impact and Advice Giving

To determine whether the journals differed in their practice of allowing advice giving or not in the abstracts of observational studies, the study focused on observational studies written by authors affiliated with institutions in the United States. The United States was chosen because it was the top publishing country in PubMed at the time the data were downloaded (Fontelo and Liu, 2018).

A generalized linear model was adopted to examine the relationship between journal impact and advice giving. This model extends linear regression models by allowing the response variable to follow distributions in the exponential family, such as normal, binomial, and Poisson distributions. Hence, the response variable can be continuous, discrete, and count (Johnson and Wichern, 2014). A main advantage of the model is that it can be used to build a regression model when the response is discrete, such as "gives-advice/does-not-give-advice" in the current analysis. Generalized linear models also include linear regression models as special cases and thus extend the applicability of the regression models.

When applying the model, the effects of relevant independent variables, including journal impact, study design, and publication year were considered. The study first identified the journal names for all the observational studies in the sample. It then used SCImago Journal Rank (SJR indicator) as a relative measure of each journal's impact. Journals with fewer than 100 papers in the dataset (Corpus-Application) were excluded from the analysis, ensuring that enough data points were included for each journal. Information os study designs and publication years was obtained from the PubMed metadata for each article.

The dependent variable was whether an article contained advice of which the value came from the BioBERT prediction results. Firstly, the study examined advice giving based on the occurrence of both weak and strong advice; then it focused on strong advice only, where authors expressed a higher level of commitment, and thus were more susceptible to quality concerns.

All the analyses were performed using the glm() function of the R package glm2 (Marschner et al., 2018). The detailed regression formula is presented in Chapter 5. For all the statistical tests, a *p*-value of < 0.001 was used to determine if there was a statistically significant difference.

3.2.4 Application Case: Retrieving Health Advice on Hydroxychloroquine Use

As there is currently no information service for direct health advice retrieval, the study further examined whether the developed model would be useful for retrieving health advice from health information services. Hence, this case study used the best-performing model to retrieve health advice from LitCovid regarding the use of HCQ to treat COVID-19. On April 30, 2021, the LitCovid corpus, which comprises 126,000 research papers, was downloaded. The MeSH ID for HCQ (MESH: D006886) was used to retrieve HCQ-related papers (Corpus-Case-Study). The prediction model was applied to all sentences in the abstracts and discussions to predict advice type.

This case study also examined whether the prediction model could be combined with current sentiment models to detect the sentiment of each advice statement, namely, if the advice is for or against the use of HCQ for COVID treatment. The sentiment analysis tool implemented to the Stanza pipeline (Qi et al., 2020) was used to get the sentiment of each advice statement. The study randomly sampled 200 advice statements, 100 with “weak advice” and 100 with “strong advice” statements. All the statements were annotated by their the advice type. Model performance was compared to the ground truth for the evaluation.

Chapter 4 Results: Part 1 – NLP Modeling of Health Advice

This chapter describes the health advice taxonomy, corpus, and prediction models. The first section provides detailed descriptions for the definitions and sentence examples for each advice category. It also tests the validity of the taxonomy by inter-coder agreement checking, as described in Chapter 3. The second section explains the process of developing the gold-standard dataset for the NLP model evaluation. The third section compares the performance of NLP-based techniques for health advice detection; the machine-learning algorithms – LinearSVM, BERT, and BioBERT – were trained and evaluated on the 6,000 annotated sentences from structured abstracts for health advice identification. The chapter also examines the model's generalizability on unstructured abstracts and discussion sections, as described in Chapter 3.

4.1 A Health Advice Taxonomy

As described in the previous section, the current study constructed a multi-dimensional taxonomy that categorized sentences in medical research abstracts and conclusions in terms of two aspects of advice: occurrence and level of commitment. By these two aspects, each sentence was categorized into "strong advice", "weak advice", or "no advice". Table 5 below shows the annotation schema and examples of sentences in each category.

Table 4: Health advice annotation taxonomy and sentence examples.

Label	Description	Example Sentence
Strong Advice	The statement makes a straightforward recommendation for health-related behavior and practice. The recommendation could lead to actionable practice and policy changes. It may target patients, health and medical professionals, or the public.	1. "Nurses ***should assess*** patient decision-making styles to ensure maximum patient involvement in the decision-making process based on personal desires regardless of age." (PMID: 26679453) 2. "A carefully integrated diabetic retinopathy

		screening service is ***needed***, particularly in remote areas, to improve adherence rates.” (PMID: 28490306)
Weak Advice	The statement hints that either a behavior or a health-related practice needs changing. Or the statement suggests that there are options and alternative approaches for a current clinical or medical practice.	3. “Adolescents with high risk factors, especially those with menstrual disorders and hyperandrogenism, ***may need*** careful clinical screening.” (PMID: 23089573) 4. “A TyG threshold of 8.5 was highly sensitive for detecting NAFLD subjects and ***may be suitable as a diagnostic criterion for*** NAFLD in Chinese adults.” (PMID: 28103934)
No Advice	The statement merely describes study background, results, findings, limitations, or calls for further research, and there is no suggestion for behavioral or clinical practice.	5. “Former smokers are at risk for hypertension, probably because of the higher prevalence of overweight and obese subjects in this group.” (PMID: 11821702) 6. “The results of the study show that in the course of HIV infection overweight/obesity affected men and women admitted with normal weight, although a greater proportion of women progressed to obesity.” (PMID: 20694301)

To test the validity of the taxonomy, a sample of 100 conclusion sentences were randomly selected for inter-coder agreement evaluation. In the current study, health advice is defined as a language construct, which does not require medical knowledge to detect. Therefore, two annotators with a background in information science and linguistic studies each labelled the 100 sentences and highlighted the linguistic cues for health advice. The overall Cohen's kappa agreement (Cohen, 1960) was 0.86, indicating a near-perfect inter-coder agreement (McHugh, 2012). Most of the disagreements occurred between "no advice" and "weak advice". Cases involving disagreements were later resolved through discussion by theannotators.

4.2 Corpus Construction

As described in Chapter 3, the developed corpus consisted of two parts: the first part was built on sentences in the structured abstracts (Corpus-Train); the second part was on sentences in unstructured abstracts and discussion sections in full-text content (Corpus-Eval).

For Corpus-Train, a total of 6,000 sentences were randomly sampled from conclusion/discussion subsections in the abstracts, including 3,000 from observational studies and 3,000 from RCTs. Based on the three-category coding schema, each sentence was assigned to one of the three category labels "no advice", "weak advice", or "strong advice". Three annotators with academic backgrounds in clinical psychology, linguistics, and information science annotated the entire training corpus. During the annotation process, they highlighted all ambiguous cases during the annotation and brought them to the team for group discussion to reach an agreement on the annotation.

In most cases, advice sentences offered only one type of advice. Occasionally, a sentence included both weak advice and strong advice. During the annotation process, these cases were treated as mixed examples and excluded from the training corpus. The final corpus contained 5,982 sentences. Since the majority of conclusion sentences did not contain advice, the category distribution in Table 6 shows a skewed distribution with "no advice" as the largest category.

Table 5: Distribution of advice typed in the annotated corpus (Corpus-Train).

	RCTs	Cross-Sectional	Case-Control	Retrospective	Prospective	Total	Percentage
None	1227	582	588	587	591	3575	59.8%
Weak	1037	82	85	144	134	1482	24.8%
Strong	652	92	45	84	52	925	15.5%
Total	2916	756	718	815	777	5982	

To evaluate the models' generalizability to sentences in unstructured abstracts and full-text content, this study randomly sampled 100 research papers (Corpus-Eval) that had unstructured abstracts and full-text access in PubMed Central (20 papers from each type of the five study designs). A total of 934 sentences from the abstracts and 3,932 sentences from the discussion/conclusion sections – which will be referred to as discussion sections for brevity – were also annotated as "strong advice", "weak advice", or "no advice". Table 7 shows the distribution of advice types in this evaluation dataset. The human annotations of this sample show that "no advice" accounted for 95.3% of the 934 unstructured abstract sentences and 92.4% of the 3,932 discussion sentences, compared to 59.8% for the conclusion subsections of the structured abstracts.

Table 6: Distribution of advice types in the unstructured abstracts and discussion sections of papers taken from Corpus-Eval.

Advice type	Unstructured abstract	Discussion section
	Total	Total
None	890	3635
Weak	28	162
Strong	16	135
Total	934	3932

4.3 Model Performance

To compare the performance of the three models, the study used macro-averaged precision, recall, and F1 scores as evaluation measures. Since the goal was to retrieve health advice, individual precision, recall, and F1 scores for each advice category were also reported. Table 8 shows the models' performance with a stratified 5-fold cross validation on the annotated corpus on sentences extracted from structured abstracts (Corpus-Train). BioBERT performed the best by all measures, achieving a macro-F1 score of 0.933. The performance of BERT was slightly lower than that of BioBERT, with a score of 0.918. This difference indicates a modest benefit of domain-specific pretraining. As both models outperformed the baseline SVM model (0.833)[2] with a wide margin, it is evident that the transformer-based method is a better choice for this task. Table 9 shows that BioBERT performed well on all kinds of advice and study designs, ranging from 0.907 to 0.943 in macro-F1 score. This indicates a low risk of prediction bias against any category.

Table 7: Model performance for detecting different types of health advice.

	Advice Type	Precision	Recall	F1
SVM	None	0.868	0.927	0.897
	Weak	0.845	0.771	0.806
	Strong	0.852	0.748	0.797
	Macro avg	0.855	0.815	0.833
BERT	None	0.949	0.943	0.946
	Weak	0.890	0.904	0.897
	Strong	0.910	0.912	0.911
	Macro avg	0.917	0.920	0.918
BioBERT	None	0.963	0.951	0.957
	Weak	0.908	0.922	0.915
	Strong	0.917	0.941	0.928
	Macro avg	0.929	0.938	0.933

Table 8:Performance of BioBERT for each study design, by F1-scores.

	RCT	Cross-Sectional	Case-Control	Retrospective	Prospective	Macro-avg
No Advice	0.919	0.971	0.983	0.973	0.966	0.955
Weak Advice	0.924	0.842	0.922	0.905	0.885	0.914
Strong Advice	0.934	0.937	0.925	0.927	0.868	0.929
Macro-avg	0.926	0.917	0.943	0.935	0.907	0.933

[2] The penalty value *C* in LinearSVM was set to 1. A comparison of different word vector representation methods showed that the tf-idf vectorization performed similarly to the count vectorization, and that adding bigrams also improved the SVM model's performance.

4.4 Error Analysis

An error analysis of misclassified cases showed that most of the prediction errors were caused by confusion between “no advice” and “weak advice”. A further examination of these errors showed that some “no advice” sentences contained confounding cues like “the importance of” or “is suitable for”, which points to implications for further study but not to health behavior changes (see example 1). Sometimes a “no advice” sentence used common advice cues such as “usefulness” and “applications” to describe study limitations instead of weak advice (see example 2), or the statement gives a vague recommendation without specifying the actions that should be taken (see example 3). There was also some confusion between “no advice” and “strong advice”. Some “no advice” sentences used strong advice cues (e.g., “is necessary”) or modal verbs (e.g., “should be”) to describe research background or implications for follow-up studies (see examples 4 and 5), and thus confused the prediction model. The following are some examples of sentences that were subject to prediction errors:

1. “Therefore, this FFQ is suitable for the investigation of nutrient-disease associations in future.”

2. “Its usefulness for this application is questionable.”

3. “Our findings could inform health policy, guide prevention strategies, and justify the design and implementation of targeted interventions.”

4. “Knowledge of molecular factors is necessary.”

5. “Further investigations should address the rationale for the early detection and control of glucose fluctuation in the era of universal statin use for CAD patients.”

4.5 Model Generalizability

4.5.1 Directly Applying the BioBERT Prediction Model

Tables 10 and 11 show the results of directly applying BioBERT, the best-performing model, to detect health advice in each of the above sentences in Corpus-Eval. The results show lower precision scores for both unstructured abstracts and discussion sections, but the recalls were comparable to those for the structured abstracts (in Corpus-Train). This means the prediction model was equally effective at retrieving health advice in unstructured abstracts and discussion sections; however, more non-advice sentences were included in the result as "false positive" predictions. Error analyses showed that these false positive predictions were mainly caused by non-advice sentences that described study background, motivation, and prior study implications. Linguistically, these sentences were very similar to advice sentences. This error pattern was actually the same as the pattern in the training data. The main reason for the increased error rate is that these confusing sentences appeared more often in unstructured abstracts and discussion sections.

Table 9: Model performance on unstructured abstracts after directly applying the fine-tuned BioBERT model.

Advice	Precision	Recall	F1	Cases
None	0.998	0.962	0.979	890
Weak	0.519	0.964	0.675	28
Strong	0.625	0.938	0.750	16
Macro avg	0.714	0.955	0.801	934

Table 10: Model performance on discussion sections after directly applying the fine-tuned BioBERT model.

Advice	Precision	Recall	F1	Cases
None	0.997	0.950	0.973	3635
Weak	0.537	0.988	0.696	162
Strong	0.696	0.881	0.778	135
Macro avg	0.743	0.940	0.815	3932

4.5.2 Improving Performance on Unstructured Abstracts

In unstructured abstracts, since health advice occurs only after a description of the results, which is near the end, a simple improvement is to assume all sentences in the first half of an abstract will not contain advice. Using this location-based filtering technique, the prediction model's precision improved to 0.900 (as shown in Table 12).

Table 11: Model performance on unstructured abstracts after applying a simple filtering rule.

Advice	Precision	Recall	F1	Cases
None	0.997	0.990	0.993	890
Weak	0.765	0.929	0.839	28
Strong	0.938	0.938	0.938	16
Macro avg	0.900	0.952	0.923	934

4.5.3 Improving Performance on Discussion Sections through Data and Feature Augmentation

Compared to that in unstructured abstracts, the distribution of health advice in discussion sections is more varied. As Figure 3 shows, although health advice, especially strong advice, tends to occurred in the second half of discussion sections, 29.3% of 297 advice sentences occur in the first half, indicating that even an optimal location filter would miss nearly a third of health advice sentences. To improve the BioBERT model's precision performance on the discussion sentences, two techniques were used: (1) augmenting the training data, and (2) adding language-style features.

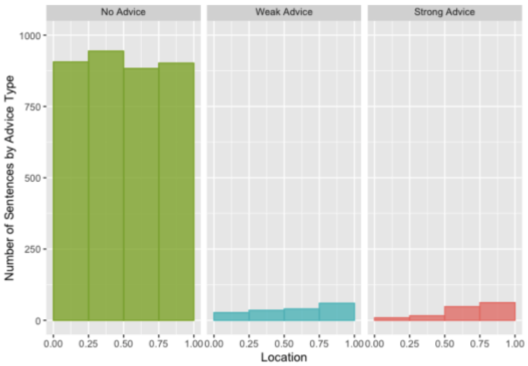


Figure 3: Distribution of health advice in discussion sections (calculated by number of sentences).

4.5.3.1 Augmenting the Training Data

The study augmented the training data (Corpus-Train) by adding the annotated discussion sentences. Overall, 3,932 discussion sentences were annotated during corpus construction, which is a considerable number of annotations. Therefore, the BioBERT model was further fine-tuned by utilizing these annotated sentences. Specifically, for each fold in the 5-fold cross-validation evaluation, 80% of the discussion sentences were added to the original 6,000 training sentences in Corpus-Train. The performance of the newly fine-tuned model was tested on the remaining 20% of the dataset.

4.5.3.2 Adding Language-Style Features

The second technique consisted of adding language-style features. It was noted that the model had not captured certain language-style markers that can effectively distinguish advice and non-advice sentences in discussion sections. The error analyses revealed two common language-style features that could confuse the model.

The first was past tense. Advice sentences do not use past tense because they used the imperative mood. In comparison, non-advice sentences can be in the past tense, despite using advice-like cues, such as “to ensure” (as shown in the sentence example below). For example:

Sentence example:

> *“We* ***took*** *great care to ensure adequate training of the neuropsychological evaluators at each site, and we monitored quality of test administration, scoring, and data entry on an ongoing basis.”*

The other common language-style marker was the citing of other studies. Advice-like sentences that contain citations are often citing advice from other studies; however, the goal of this study was to identify advice given by the authors in the current study rather than advice given in prior studies. The sentence example below shows how authors cite the advice in prior studies.

Sentence example:

> *"NMDA receptor antagonists such as ketamine or magnesium have been suggested for postoperative pain management* ***[22,23]****."*

To add the language-style markers into the model, the current work augmented the BioBERT input (a single sentence) was augmented with the following three "binary" features: (1) data source: whether a sentence is from a structured abstract or a discussion section, (2) citation: whether a sentence contains a citation, and (3) past tense: whether a sentence uses past tense.

When integrating the above features into the BioBERT model, the special BERT mark [SEP] was used to concatenate the features with the original sentence via the format below:

data source [SEP] citation [SEP] past tense [SEP] sentence

For example, a sentence from a discussion section that used past tense but did not cite other studies was represented as:

discussion [SEP] No [SEP] Yes [SEP] sentence

All the sentences from structured abstracts were represented in the following form:

structured abstract [SEP] [SEP] [SEP] sentence

Tables 13-16 show that augmenting the training data resulted in a significant improvement in the macro-F1 score to 0.864. The added language-style features further improved the F1 score to 0.907.

Table 12: Model performance on discussion sections after fine-tuning the BioBERT model (data augmentation only).

Advice	Precision	Recall	F1	Cases
None	0.991	0.977	0.984	3635
Weak	0.708	0.883	0.786	162
Strong	0.793	0.852	0.821	135
Macro avg	0.831	0.904	0.864	3932

Table 13: Model performance on discussion sections after fine-tuning the BioBERT model (data augmentation + data source).

Advice	Precision	Recall	F1	Cases
None	0.987	0.987	0.987	3635
Weak	0.781	0.815	0.798	162
Strong	0.875	0.830	0.852	135
Macro avg	0.881	0.877	0.879	3932

Table 14: Model performance on discussion sections after fine-tuning the BioBERT model (data augmentation + data source + has citation).

Advice	Precision	Recall	F1	Cases
None	0.989	0.986	0.988	3635
Weak	0.806	0.846	0.825	162
Strong	0.833	0.852	0.842	135
Macro avg	0.876	0.895	0.885	3932

Table 15: Model performance on discussion sections after fine-tuning the BioBERT model (data augmentation + data source + has citation + past tense).

Advice	Precision	Recall	F1	Cases
None	0.991	0.990	0.990	3635
Weak	0.827	0.883	0.854	162
Strong	0.892	0.859	0.875	135
Macro avg	0.903	0.911	0.907	3932

4.6 Summary

This chapter described the health advice taxonomy and annotated corpus developed in this study. It also compared the performance of both traditional machine-learning and deep-learning approaches for health advice detection. The results show that the developed BioBERT-based model outperformed the BERT-based model and the SVM. The high performance of the BioBERT model on all measures suggests that this transformer-based deep-learning approach is a better choice of the task. The better performance of BioBERT over BERT's also indicates that the study benefited from domain-adapted pre-training. The generalizability evaluation shows that with some tuning, the developed BERT-based model on structured abstracts is able to be generalized well to sentences in the unstructured abstracts and the discussion sections of papers for advice statement detection.

Chapter 5 Results: Part 2 – Examining Health Advice Giving Behavior in Medical Literature

Based on the corpus and prediction models relating to RQ1, this chapter further presents the experimental results of the study. First, it examines whether advice statements from both abstracts and discussion sections are equivalent, by comparing differences in their level of commitment and the amount of information (as described in Chapter 3.2.1). Second, it measures the relationship between journal impact and advice giving, with the purpose of examining whether journals differ in the practice of allowing advice giving in medical research papers (as described in Chapter 3.2.2). Third, it describes the health advice that has been made regarding the use of HCQ for COVID-19 treatment. It then evaluates whether current sentiment analysis tools can be combined with the current model to identify polarities in each advice statements (as described in Chapter 3.2.3).

5.1 Advice Giving in Abstracts and Discussion Sections

Health advice can occur in both abstracts and discussions. Thus, if a research paper gives equivalent advice statements in both sections, a health advice detection service needs to retrieve advice from abstracts only. Otherwise, access to full-text content is needed to navigate and summarize health advice.

Based on the annotations of 100 medical research papers (Corpus-Eval with both abstracts and discussions in Chapter 4), the study aggregated the sentence-level annotations into section-level values and compared researchers' advice-giving behavior in the abstracts and discussions. To answer RQ3, it compared (1) the numbers of papers that gave advice with those that did not and (2) the strength of the advice made in the abstracts versus the discussion sections.

5.1.1 Advice Aggregation and Comparison

For the analysis, the study aggregated the sentence-level labeling result to the section level, using the rule that a section contained health advice if at least one abstract or conclusion/discussion sentence did so. Furthermore, if both weak and strong recommendations were found, the "strong advice" label was applied to the section. The section-level labels made it possible to count the number of advice-giving articles and to analyze the practice of giving health advice.

To compare the health advice in the abstract and the discussion section of each paper, the study assigned each paper to one of the following four groups:

(1) group 1 (the "advice-in-neither-section" group) included articles that gave no advice in their the abstract or the discussion section;

(2) group 2 (the "advice-in-abstract-only" group) included articles that gave advice only in the abstract;

(3) group 3 (the "advice-in-discussion-only" group) included articles that gave advice only in the discussion section;

(4) group 4 (the "advice-in-both-sections" group) included articles that gave advice in both the abstract and the discussion section.

As for the strength of health advice in the abstracts and the discussion sections, the study compared two aspects of each article: (1) whether the recommendations made in the abstracts and discussion sections were semantically similar and discussed the same clinical or policy practice, and (2) in the case of semantically similar recommendations, whether the abstracts or discussion sections gave stronger advice. Figure 4 below illustrates our process of analysis.

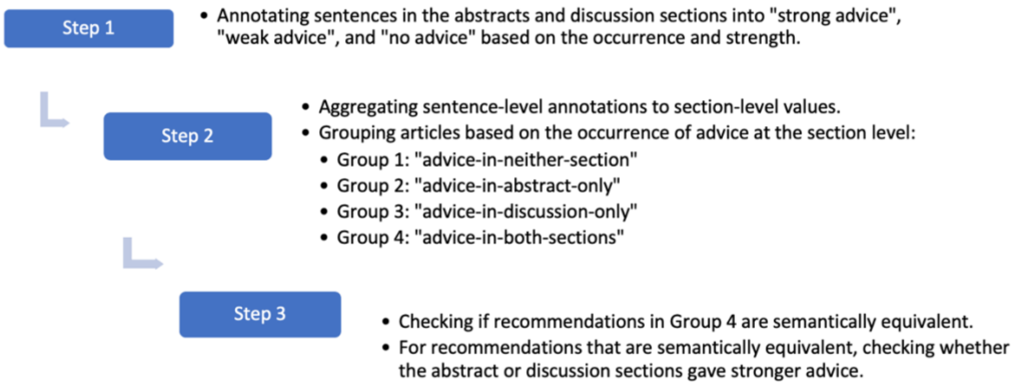

Step 1
Annotating sentences in the abstracts and discussion sections into "strong advice", "weak advice", and "no advice" based on the occurrence and strength.
Step 2
Aggregating sentence-level annotations to section-level values.
Grouping articles based on the occurrence of advice at the section level:
Group 1: "advice-in-neither-section"
Group 2: "advice-in-abstract-only"
Group 3: "advice-in-discussion-only"
Group 4: "advice-in-both-sections"
Step 3
Checking if recommendations in Group 4 are semantically equivalent.
For recommendations that are semantically equivalent, checking whether the abstract or discussion sections gave stronger advice.

Figure 4: An illustration of the analysis process.

5.1.2 Advice Giving in Abstracts and Discussion Sections

Table 17 shows the category distribution of sentences in the abstracts and discussion sections. The average length of the abstracts was nine sentences. The average length of the discussions was 39 sentences. Overall, the discussion sections contained a higher percentage of advice sentences (7.6% total, 3.5% strong, 4.1% weak), compared to that in the abstracts (4.7% total, 1.7% strong, 3.0% weak). To count the articles that gave advice and to compare the strength of advice between the abstracts and discussion sections, the sentence-level annotations were aggregated into section-level values.

Table 16: Advice-sentence distribution in the abstracts and discussion sections.

Advice Type	Abstract	Discussion
No advice	890 (95.3%)	3635 (92.4%)
Weak advice	28 (3.0%)	162 (4.1%)
Strong advice	16 (1.7%)	135 (3.5%)
Total	934	3932

Table 18 shows the distribution of health advice after the sentence-level annotations were aggregated at the section level. The results show that only 20% of articles gave no advice in either the abstract or discussion, suggesting that a majority of researchers embrace the practice of giving health advice in individual studies. However, only 2% (2/100) of articles gave advice in abstracts only, which means that smost papers (78%) did not use the abstract as the main place to give advice. In fact, nearly half the papers (45%) chose to give advice in the discussion section only, and 33% gave advice in both the abstact and the docussion section. Overall, researchers were much more likely to give advice in the discussion section than in the abstract, and they rarely gave advice only in the abstract.

Table 17: Distribution of articles based on the practice of advice giving.

Article Type	Counts	Percentage
Advice-in-neither-section	20	20.0%
Advice-in-abstract-only	2	2.0%
Advice-in-discussion-only	45	45.0%
Advice-in-both-sections	33	33.0%
Total	100	

Table 19 shows the advice types for the 33 articles that gave advice in both the abstract and the discussion: nine gave advice in both sections; 12 gave strong advice in both sections; and 12 gave strong advice in the discussion but weak advice in the abstract. Interestingly, none of the papers gave strong advice in the abstract and weak advice in the discussion. These results suggest that the researchers were cautious about giving advice, especially about giving strong advice in an abstract.

Table 18: Distribution of articles giving advice in both the abstract and the discussion.

		Discussion	
		Weak Advice	Strong Advice
Abstract	Weak Advice	9 (27%)	12 (36%)
	Strong Advice	0	12 (36%)

To compare the content of different advice sentences in a particular article, this chapter further checked the correspondence between the advice in the abstract and the discussion. The study paid particular attention to whether an author offered multiple versions of the same recommendations with inconsistent level of commitment. The results show no strength inconsistency in the 12 articles that gave weak advice in abstracts and strong advice in discussions. Instead, two strategies were identified which the authors used to give different versions of advice. One strategy was to give weak and non-specific advice in the abstract, while using more sentences to give a completely different version of advice, stronger and more specific, in the discussion section. This strategy occurred in five of the 12 articles. In the following examples, the author gives a weak and non-specific recommendation for a treatment protocol that is useful (sentence 1 in the abstract). In comparison, the author makes a series of direct recommendations for a specific clinical practice, adding a number of conditions required for their implementation (sentences 1-4 in the discussion section).

PMC: 5808411

Section: Abstract

1. "Thus, a protocol for clinicians to manage the patient presenting with oligometastatic prostate cancer ***would be a useful clinical tool****."*

Label: weak advice

Section: Discussion

1. "As in other settings, only those patients likely to suffer mortality or substantial morbidity due to their ***disease should be considered*** *for aggressive treatment****, which should only be offered*** *in the setting of an institutional-review-board-approved clinical trial or prospective registry."*

Label: strong advice

2. "Patients ***must be fully informed*** *of the potential risks and benefits associated with an aggressive approach; specifically, they* ***must be made aware*** *that data from appropriately conducted studies to demonstrate prolonged survival as a result of treatment is lacking."*

Label: strong advice

3. "Men who do undergo treatment ***should be assessed and treated*** *in a multidisciplinary setting including medical oncology, radiation oncology, and urology."*

Label: strong advice

4. "Clinicians managing such patients ***should consider*** *establishing a prostate cancer multidisciplinary clinic if not already present at their institution."*

Label: strong advice

5. "Finally, establishment of an institutional biorepository for banking of serum, urine, stool, and tissue samples ***should be considered*** *– only with the committed and coordinated efforts of the entire health-care team will we find answers to the many questions that remain."*

Label: strong advice

The other strategy was to use more sentences in the discussion for stronger and more specific recommendations but to include two paraphrased, but semantically equivalent, sentences in the abstract and the discussion section (this was the case in 7 of the 12 articles). The sentence examples below show a pair of semantically similar recommendations extracted from an abstract and a discussion section.

PMC: 325258

Section: Abstract

"Therefore, intraoperative antifibrinolysis ***may not be indicated in routine cardiac surgery*** *when other blood-saving techniques are adopted."*

Label: weak advice

Section: Discussion

"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****."*

Label: weak advice

Overall, when giving advice in both abstracts and discussions, the researchers tended to give weak and non-specific advice in the abstract, while giving stronger and more specific advice in the discussion section, where there is more room to lay out the conditions required for the strong recommendations.

5.1.3 Summary

This section has presented the results from a comparison of health advice in abstracts and discussion sections. The findings show that health researchers commonly give advice in individual studies; however, they rarely give advice only in the abstract. It is more common for them to give advice only in the discussion section, or in both the abstract and the discussion. When giving advice in both sections, researchers tend to give weak and non-specific advice in the abstract, usually in one sentence, and to give strong and more specific advice in the discussion section, using more sentences and describing the conditions required to implement the recommendations. The results suggest that most researchers support giving advice in individual studies but that they are generally cautious about giving advice in abstracts.

5.2 Health Advice in Observational Studies

Prior studies of health advice argue that the prevalence of health advice in individual studies, especially observational studies, may raise concerns about scientific rigor, since clinical recommendations in observational studies sometimes make a substantial logical leap without evidential support from the study. However, the prior estimations were based on manual content analyses of small samples of articles, which is not adequate for judging the severity of the problem over the years. In this subsection, the health advice prediction model is applied to estimate the prevalence of health advice in observational studies on a large scale (Corpus-Application). In addition, it measures the relationship between journal impact and advice giving.

5.2.1 Advice Prediction and Aggregation

The dataset that was downloaded (Corpus-Application) included 1,620,870 conclusion sentences from 832,671 observational studies with structured abstracts. The study used BioBERT, the best-performing model, to identify the health advice in each sentence. To evaluate the models' generalizability to all observational studies, 100 sentences in the prediction results were chosen by random sampling, and their accuracy was manually checked. The human annotations of this sample showed that "no advice" accounted for 70% of the examples, followed by "weak advice" (17%) and "strong advice" (13%). The results show scores of .86 for precision, .89 for recall, .87 for macro-F1, and .90 for accuracy. This accuracy level is slightly lower than the cross-validation results on the training corpus (as presented in Chapter 4). However, the major type of error was still a failure to distinguish between "no advice", and "strong/weak advice". Overall, the model generalized well to all the observational studies, although the higher number of "no advice" sentences may have presented a challenges to the prediction model.

The sentence-level predictions were then aggregated to article-level predictions. Specifically, an article was considered to contain health advice, if at least one sentence in the conclusion subsection had health advice. An article was considered as containing weak advice, if at least one sentence had weak health advice and no sentence had strong advice. An article was rated as having strong advice, if at least one sentence had strong advice.

5.2.2 The Prevalence of, and Trends in, Advice Giving

Based on the article-level predictions, the study counted the number of advice-giving articles in each study design group (see Table 20 for the distribution). Among the 832,671 observational studies, 342,973 (41.2%) contained health advice: 187,275 (22.5%) contained only weak advice, and 155,698 (18.7%) contained strong advice. This estimation is much lower than the 56% estimated by Prasad et al. (2013) using manual content analysis on a small sample.

Table 19: Distribution of health advice in the prediction results for the observational studies (at the article level).

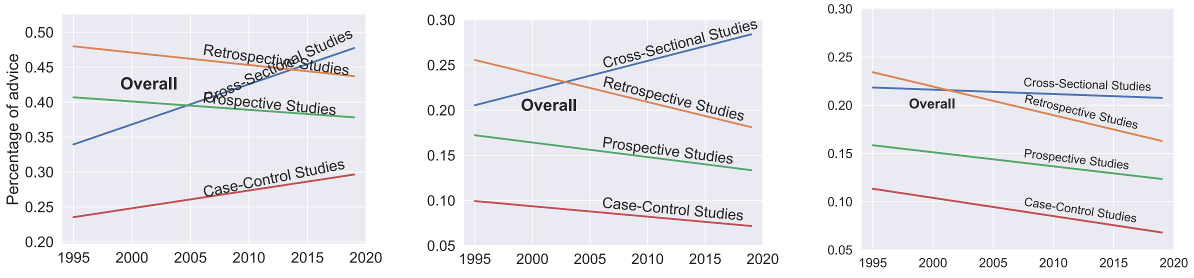
	Cross-Sectional	Case-Control	Retrospective	Prospective	Total
No Advice	82,113	68,864	221,903	116,818	489,698
Weak Advice	25,363	17,980	98,209	45,723	187,275
Strong Advice	37,261	7,593	82,595	28,249	155,698
Total	144,737	94,437	402,707	190,790	832,671

Figure 5a plots the ratios of the observational studies that have provided health advice in the past 25 years, from 1995 to 2019. The studies were grouped by design type, and the trend in each group was examined. Figure 5a shows that although the overall trend to provide health advice has been increasing over the past 25 years, trends in the different study design groups are inconsistent: the health advice ratios in retrospective and prospective studies have slightly decreased, while the ratios in case-control and cross-sectional studies have increased. The ratio has increased the most in cross-sectional studies, from a low of 34% in 1996 to a high of 51% in 2019.

In addition to comparing trends in the different study groups, pattens in the “strong advice” group were further examined. In these groups, the authors expressed higher levels of commitment, and hence, the advice was more susceptible to inaccuracies. Figure 5b illustrates the overall trend in the giving of “strong advice”. Compared to the overall trend in giving advice, the ratio of “strong advice” has fluctuated between 18% and 21% without a significant change over the past 25 years. More surprising, the ratio of “strong advice” in case-control studies has decreased over the years, in contrast to the increasing trend in giving “weak” and/or “strong” advice, as shown in Figure 5a.

This difference indicates that the increase in giving "weak advice", not "strong advice", drives the upward trend in case-control studies. In observational studies authored by researchers from the United States, the results in Figure 5c show no upward trend in any of the four study design groups; the trend is nearly flat in cross-sectional studies and decreases in the other types of study.

The patterns in Figures 5a, b, and c draw a different and more complicated picture than that reported in previous studies such as Prasad et al. (2013). Based on the observations made of this dataset, previous concerns about the prevalence of health advice in observational studies might be overdone. The decreasing trend in some study design groups (i.e., retrospective and prospective studies) and regions (i.e., the United States) suggests that the research community may have become more rigorous in vetting health advice in observational studies over the past few decades. These results also provide evidence for the claim that advice giving in observational studies varies across different study designs and countries, and therefore challenge the notion of a shared consensus on whether to give health advice, either weak or strong, in science communication.



(a) Globally (all countries) (b) Globally (strong advice only) (c) U.S. only (strong advice only)

Figure 5: Trends in the giving of health advice in the four types of observational study designs (1995-2019).

5.2.3 Journal Impact and Advice Giving

Using the PubMed metadata, 3,911 journals were identified as the publication venues for the 832,671 observational studies labeled by PubMed. The study then used the SCImago Journal Rank (SJR indicator) as the measure for journal impact. The study further preprocessed the journal impact values with log transformation, due to their highly skewed distribution. Journals with less than 100 papers in the data set were excluded from the analysis. The final dataset included 402 journals and 154,339 papers authored by researchers from the United States.

Table 21 gives the formula that was used in the analysis to construct the regression model with the glm( ) function in the R package lgm2.

Table 20: Formular for the logistic linear regression model.

advice giving ~	
journal impact	// numerical
+ study design	// cross-section, case-control, retrospective, prospective
+ year	// numerical

Applying the generalized linear model for regression analysis, the study found a significant journal impact difference on advice giving when other factors were controlled for, including study design and publication year (for both weak and strong advice). A negative association between the journals' log-scaled impact factors and the health advice ratios was found (coefficient = -0.32, standard error = 0.01, z value = -24.34, *p-value*≪0.0001). This suggests that observational studies in low-impact journals are more likely to contain health advice.

As for strong advice, the study also observed a negative association between the journals' log-scaled impact factors and advice ratios (coefficient = -0.16, standard error = 0.02, z value = -9.28, *p-value*≪0.0001). This result is consistent with the claim in the previous manual content analyses that higher-impact journals are less likely to contain health advice (Wilson and Chestnutt, 2016).

5.2.4 Summary

This section has presented the results on the prevalence of advice giving and on related trends in observational studies over the past 25 years. The findings from the analysis show that although health advice consistently appeared in different types of observational studies, the overall ratio was lower than that reported by prior content analyses. Furthermore, journals with lower impact factors are more likely to include health advice in the abstracts of observational studies by researchers from the United States than journals with higher impact factors. The differences among subgroups in respect to health advice in observational studies call for further fine-grained analyses.

5.3 Health Advice on HCQ Use for COVID-19 Treatment

This section presents the results on retrieving health advice about HCQ as a treatment for COVID-19. It is a case application of the developed NLP technique for advice detection, as described in chapter 3.2.3. The section first presents the health advice that has been offered regarding the use of HCQ for COVID-19 treatment. It then evaluates whether current sentiment analysis tools can be combined with the current model to further identify polarities within advice statements.

5.3.1 Health Advice on HCQ Use

LitCovid organizes all medical research papers by topic, including “transmission”, “diagnoisis”, “prevention”, and “treatment”. It also tags all chemicals that have been studied and reported on in research papers. Using the MeSH ID D006886 for HCQ, the study retrieved 3,400 HCQ-related papers from the 126,000 research papers in LitCovid. Among the related papers, 10,000 sentences tagged with HCQ or its alternative names, such as *hydroxychloroquine*, and *(hydroxy)chloroquine sulfate* are retrieved. These sentences were then sent to the trained BioBERT model to identify HCQ-related health advice.

In the prediction results, this study found 605 strong advice statements and 815 weak ones. Via content analysis, the study noticed that the detected strong and weak advice statements mainly fell into the following four categories:

(1) recommendations for using HCQ to treat COVID-19:

Sentence example:

"We therefore recommend that COVID-19 patients be treated with hydroxychloroquine and azithromycin to cure their infection and to limit the transmission of the virus to other people in order to curb the spread of COVID-19 in the world."

(2) recommendations on the dosages and use of HCQ

Sentence example:

"In order to meet predefined HCQ exposure target, HCQ dose may need to be reduced in young children, elderly subjects with organ impairment and/or coadministration with a strong CYP2C8/CYP2D6/CYP3A4 inhibitor, and be increased in pregnant women."

(3) cautions and warnings about HCQ use

Sentence example:

"Additionally, hypoglycemia must be looked for in patients with diabetes especially with concurrent use of chloroquine/HCQ and lopinavir/ritonavir."

(4) recommendations not to use HCQ to treat COVID-19:

Sentence example:

"Taken together, HCQ should not be used in prophylaxis against COVID-19."

The case study demonstrates that this health advice prediction model can be combined with current health information service systems to provide more convenient navigation of a large volume of health literature.

5.3.2 Polarities in HCQ-Related Advice

For the evaluation of the polarity analysis tool, 100 weak and 100 strong advice statements were randomly sampled from the prediction results. Each advice statement was manually labelled as "positive", "negative", or "neutral" to indicate the polarity of its stance towards the use of HCQ.

The “positive” statements supported the use of HCQ to treat COVID-19 The “negative” class objected to its use. The “neutral” class was neither positive nor negative. Table 22 below shows the distribution of advice polarities for the weak and strong advice statements.

Table 21: Distribution of the annotated polarities of HCQ-related health advice statements.

	Weak advice	Strong advice	Total
Positive	36	23	59
Negative	16	9	25
Neutral	48	68	116
Total	100	100	200

Table 23 presents the confusion matrix for the detected polarities by the Stanza sentiment analysis tool and the annotated polarity. Table 24 further shows the precision, recall, and F1-scores of the prediction results for the 200 randomly sampled advice statements.

Table 22: Confusion matrix comparing the annotated and predicted polarities of health advice.

		Prediction		
		Positive	Negative	Neutral
Annotation	Positive	15	33	11
	Negative	0	24	1
	Neutral	14	87	15

Table 23: Performance of the Stanza sentiment analysis tool for detecting advice polarity.

Polarity	Precision	Recall	F1 Score	Cases
Positive	0.517	0.254	0.341	59
Negative	0.167	0.960	0.284	25
Neutral	0.556	0.129	0.210	116
Macro avg	0.413	0.448	0.278	200

The evaluation results show that the sentiment analysis tool had difficulty distinguishing positive advice sentences from negative ones. In fact, many of the positive and neutral examples were wrongly assigned to the negative class. Error analyses showed that most of errors were caused by confounding cues in health advice. For example, words such as *infection* and *warning*, as shown in sentence examples 1 and 2 below, do not indicate any polarity towards HCQ use, but they may trick the model into classifying them as negative.

Sentiment examples:

1. "Our results foster ***warnings*** *before initiating a treatment with HCQ in patients, regardless of its indication."* (Annotation: positive; prediction: negative.)

2. "It advises that hydroxychloroquine/chloroquine should be continued for SLE, even in the context of active COVID-19 ***infection****."* (Annotation: positive; prediction: negative.)

In addition, researchers normally use cues such as *advocate* or *recommend* to support the use of HCQ (as shown in sentence example 3 below). However, these indicators are not necessarily the same as those in the review datasets commonly used to train sentiment analysis tools. Therefore, unless similar polarity indicators such as *good, not, against* are used for or against the use of HCQ in health advice (as shown in examples 4-6), the model could fail to identify the polarity correctly.

3. "We ***advocate*** *the use of hydroxychloroquine in the management of type 2 lepra reactions as a steroid sparing agent as per recent Govt."* (Annotation: positive; prediction: neutral.)

4. "Our findings do ***not*** *support the routine use of azithromycin in combination with hydroxychloroquine in patients with severe COVID-19."* (Annotation: negative; prediction: negative.)

5. "Taken together, HCQ ***should not*** *be used in prophylaxis against COVID-19."* (Annotation: negative; prediction: negative.)

6. "However, these possible side effects of hydroxychloroquine plus the negative clinical results of this study argue ***against*** *the widespread use of hydroxychloroquine in patients with covid-19 pneumonia."* (Annotation: negative; prediction: negative.)

The results suggest that current sentiment analysis tools such as the Stanza are inadequate for advice polarity detection. Other NLP tools such as claim and stance classification tools (e.g., Anand et al., 2012; Ferreira and Vlachos, 2016; Li et al., 2017; Yu et al., 2019; Kilicoglu et al., 2019), may further aggregate health advice regarding HCQ. None of these functions are available in current health information services like LitCovid; however, based on the health advice detection model, they could be built to benefit health researchers and practitioners in the future.

Chapter 6 Discussion and Conclusion

This chapter summarizes the main results and findings obtained in this study and discusses ethical implications and possible future research to investigate questions arising from this dissertation work.

6.1 Discussion

For this study, an annotation taxonomy and a corpus of health advice were developed based on the occurrence of health advice and its level of commitment, as well as NLP-based techniques for automatically identifying health advice in the medical research literature. As a valuable source for health advice detection and analysis, the dataset and code are publicly available, which can have a significant impact on science communication and education.[3] The study revealed that the BioBERT model outperformed BERT and SVM on the training dataset developed for advice classification. This result suggests that, compared to the traditional machine-learning model, a transformer-based method is a better choice for the task. The better performance of the BioBERT over the BERT model also suggests there is a modest benefit of domain-specific pretraining for advice detection in the medical research literature.

The content analysis of health advice in abstracts and discussion sections show that most researchers support the giving of advice in individual studies but that they are generally cautious about giving advice in abstracts. The common practice is that researchers rarely give health advice only in an abstract. When advice statements are given in both the abstract and discussion section of a paper, researchers tend to give advice with a higher level of commitment. The findings from the current analysis also indicate that health advice in abstracts, although widely accessible, does

[3] The annotated corpus and code are available at: https://github.com/junwang4/detecting-health-advice

not contain details, such as the specific instructions for implementing a clinical practice. Therefore, readers of medical research publications should check the discussion sections in the full text of an article for a thorough review of the implications. This finding also calls for open access to medical research publications, so that the clinical and policy recommendations can be understood accurately by health professionals and the general public.

The prevalence and trend analyses of health advice in the sample of observational studies showed that of the 832,671 observational studies, 342,973 studies (41.2%) contained health advice: 187,275 (22.5%) contained weak advice only, and 155,698 (18.7%) contained strong advice. This estimation of articles containing health advice is much lower than the 56% estimation by Prasad et al. (2013), which was arrived through manual content analyses of a small sample. Although the overall trend in advice giving has increased over the past 25 years, the trends in different study design groups are inconsistent: the health advice ratios in retrospective and prospective studies are slightly decreasing, while the ratios in case-control and cross-sectional studies are increasing.

The regression analysis of journal impact, using the generalized linear model, showed that articles published in lower-impact journals were more likely to give health advice, which consisted of both weak and strong advice, and strong advice only for authors from the United States. The relationship between journal impact and advice giving provides evidence that journals are important gatekeepers for vetting the health advice offered by authors. As health advice in abstracts can have a greater reach and impact, the results of this study call for verification of the health advice in the medical research literature, especially advice given in abstracts.

The case application of the NLP model developed for retrieving health advice from LitCovid on the use of HCQ for COVID-19 treatment indicates that the model may be used with current information services to summarize health advice on specific health topics. The results from the evaluation of advice polarity detection show that current sentiment analysis tools might not be directly applicable to the developed model to summarize advice polarities. However, the better recall scores for advice with a negative polarity suggest that the existing sentiment analysis tools should be able to detect advice polarity if they are provided with clear sentiment indicators.

In summary, this study involved the development of a high-performing NLP model that can detect weak and strong health advice in abstracts and discussion sections in medical research publications. The case application of the model to research papers in LitCovid also demonstrated that it may be combined with health information services to navigate and summarize health advice in a large number of research outputs. Health researchers, practitioners, and the public could also use the model to track health advice in individual studies. By linking the collected advice with indications of the strength of evidence, and by relying on domain expertise, the developed model could also help to verify the quality of advice, supplementing and augmenting human intelligence in health information assessment.

Applying the developed corpus and model to research papers in PubMed, this study also revealed a relationship between journal impact and advice giving, which highlights journals' gatekeeping role in allowing the giving of health advice in medical publications. It also calls for cautions from health researchers, practitioners, and the public in adopting health advice from individual studies.

It should be noted that although the strength of advice (i.e., level of commitment) is arbitrarily defined based on the strength of linguistic cues, the interpretation of advice strength may differ based on the topic. For example, advice about urgent topics such as COVID-19 may be stronger than advice on other topics, such as chronicle diseases.

6.2 Ethical Implications

The following ethical issues are relevant to this study:

1. The NLP model developed here was designed to identify sentences that provide health advice in medical papers. However, this model cannot verify whether an instance of health advice is valid or not.

2. As discussed in the introduction section and the case application, health advice given in individual research papers may fail to provide sufficient evidence or may be outdated; hence verification by health professionals is called for before the advice is implemented in clinical use.

3. Researchers often write for professional audiences, and thus, they may have provided health advice intended for health professionals instead of the general public. Furthermore, the interpretation of health advice may also require more context than a sentence or two. Therefore, average users are urged to discuss with their doctors whether they should follow health advice found by this NLP model.

4. For the same reason, for times when this model is used in real-world situations, the application developers should provide a function that flags or removes inaccurate or outdated health advice upon request from authors and health experts.

5. Although this NLP model achieves a high level of prediction accuracy, false positive and false negative predictions may still occur. While false positive predictions (non-advice sentences in the result) may just be a nuisance, false negative predictions (missed health advice) may cause misunderstandings if the model is being used to retrieve all health advice on a topic.

6. Users should be trained to understand that the model does not provide a perfect recall for retrieving all the advice in the medical research literature.

6.3 Future Research Directions

The experimental results of this study might have raised more questions than they answered. This section describes new research problems raised by this work and suggests possible future research toward for solving these problems.

6.3.1. Fine-Grained Health Advice Analysis

The advice prediction model focuses mainly on detecting advice and its level of commitment. For better advice retrieval and summarization in real practice, more fine-grained health advice analysis is needed. For example, computational models could be developed to identify health advice that raises awareness and advice that includes suggestions for clinical treatment. Identifying more types of advice could help information seekers to navigate the advice they are seeking for.

In the meantime, linguistic frameworks such as claim specificity (i.e., how much detailed information is included) could be adopted for synthesizing and aggregating medical advice. A multi-dimensional advice taxonomy and annotation corpus could be developed to describe advice statements based on their aspects (e.g., “clinical practice”, “policy change”), stances (e.g., “support”, “object to”, “neutral”), and target audience (e.g., “patients”, “practitioners”, “policymakers”). Similary, computational models could be developed to track advice or scientific implications with finer granularity and facilitate downstream applications such as advice sentiment, detection and advice quality verification.

In addition to journal impact, factors such as authors’ institution type, career stages might also affect their advice-giving behavior. The is could be work for future studies.

6.3.3 Health Advice in News and Social Media

The misrepresentation of scientific studies and implications of their findings, such as health advice reported in science news and social media, has been identified as a major problem in science communication. This dissertation focuses on advice in the medical research literature. However, the model could be further tuned and evaluated to detect health advice on other information subsides, such as news outlets and social media. A future direction is to apply the model to extract health advice from different resources. By extracting and comparing different versions of health advice, we should be able to detect overclaimed or underclaimed advice during the diffusion process and identify the factors related to advice distortion.

References

[1] Adel, H., & Schütze, H. (2017, April). Exploring Different Dimensions of Attention for Uncertainty Detection. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers* (pp. 22–34).

[2] Agarwal, S., & Yu, H. (2010). Biomedical negation scope detection with conditional random fields. *Journal of the American Medical Informatics Association: JAMIA*, *17*(6), 696.

[3] Alamri, A., & Stevenson, M. (2016). A corpus of potentially contradictory research claims from cardiovascular research abstracts. *Journal of Biomedical Semantics*, *7*(1), 36.

[4] Alamri, A. (2016). *The Detection of Contradictory Claims in Biomedical Abstracts* (Doctoral dissertation, University of Sheffield).

[5] Anand, P., Walker, M., Abbott, R., Tree, J. E. F., Bowmani, R., & Minor, M. (2011, June). Cats rule and dogs drool!: Classifying stance in online debate. In *Proceedings of the 2nd Workshop on Computational Approaches to Subjectivity and Sentiment Analysis* (pp. 1–9). Association for Computational Linguistics.

[6] Andrews, J., Guyatt, G., Oxman, A. D., Alderson, P., Dahm, P., Falck–Ytter, Y., Nasser, M., Meerpohl, J., Post, P.N., Kunz, R., & Brozek, J. (2013). GRADE guidelines: 14. Going from evidence to recommendations: the significance and presentation of recommendations. *Journal of Clinical Epidemiology*, *66*(7), 719–725.

[7] Anderson, L. B. (1986). Chafe, W. & Nichols, J. (Eds.), *Evidentiality: The linguistic coding of epistemology. Advances in discourse processes*, Ablex Publishing, Norwood, NJ (1986), pp. 273–312

[8] Austin, J. L. (1975). *How to do things with words* (Vol. 88). Oxford University Press.

[9] Banerjee, R., & Prasad, V. (2020). Are observational, real–world studies suitable to make cancer treatment recommendations?. *JAMA Network Open*, *3*(7), e2012119.

[10] Barton, E. L. (1993). Evidentials, argumentation, and epistemological stance. *College English*, *55*(7), 745–769.

[11] Bavelas, J. B., Black, A., Chovil, N., & Mullett, J. (1990). Truths, lies, and equivocations: The effects of conflicting goals on discourse. *Journal of Language and Social Psychology*, *9*(1–2), 135–161.

[12] Bekker, H. L., Winterbottom, A. E., Butow, P., Dillard, A. J., Feldman–Stewart, D., Fowler, F. J., Jibaja–Weiss, M.L., Shaffer, V.A. and Volk, R.J., & Volk, R. J. (2013). Do personal stories make patient decision aids more effective? A critical review of theory and evidence. *BMC Medical Informatics and Decision Making*, *13*(S2), S9.

[13] Bies, A., Mott, J., Warner, C., & Kulick, S. (2012). English web treebank. *Linguistic Data Consortium, Philadelphia, PA*. https://doi.org/10.35111/m5b6–4m82

[14] Boutron, I., Dutton, S., Ravaud, P., & Altman, D. G. (2010). Reporting and interpretation of randomized controlled trials with statistically nonsignificant results for primary outcomes. *JAMA*, *303*(20), 2058–2064.

[15] Boutron, I., & Ravaud, P. (2018). Misrepresentation and distortion of research in biomedical literature. *Proceedings of the National Academy of Sciences*, *115*(11), 2613–2619.

[16] Brainard, J. (2020). Scientists are drowning in COVID–19 papers. Can new tools keep them afloat. *Science*, *13*(10), 1126.

[17] Brown, P., & Levinson, S. (1987). *Politeness: Some language universals in language use*. Cambridge: Cambridge University Press.

[18] Brown, P., Brunnhuber, K., Chalkidou, K., Chalmers, I., Clarke, M., Fenton, M., Forbes, C., Glanville, J., Hicks, N.J., Moody, J., & Twaddle, S. (2006). How to formulate research recommendations. *BMJ, 333*(7572), 804–806.

[19] Brun, C., & Hagege, C. (2013). Suggestion mining: Detecting suggestions for improvement in users' comments. *Research in Computing Science.*, *70*(79), 171–181.

[20] Butow, P. N., Kirsten, L. T., Ussher, J. M., Wain, G. V., Sandoval, M., Hobbs, K. M., Hodgkinson, K. & Stenlake, A. (2007). What is the ideal support group? Views of Australian people with cancer and their carers. *Psycho-Oncology: Journal of the Psychological, Social and Behavioral Dimensions of Cancer*, *16*(11), 1039–1045.

[21] Bybee, J., & Fleischman, S. (1995). Modality in grammar and discourse: An introductory essay. *Modality in Grammar and Discourse*, *14*, 503–517.

[22] Cabanski, T. (2019, June). DS at SemEval–2019 Task 9: From Suggestion Mining with neural networks to adversarial cross–domain classification. In *Proceedings of the 13th International Workshop on Semantic Evaluation* (pp. 1192–1198).

[23] Chang, C. (2015). Inaccuracy in health research news: A typology and predictions of scientists' perceptions of the accuracy of research news. *Journal of Health Communication*, *20*(2), 177–186.

Chapman, W. W., Bridewell, W., Hanbury, P., Cooper, G. F., & Buchanan, B. G. (2001). A simple algorithm for identifying negated findings and diseases in discharge summaries. *Journal of Biomedical Informatics*, *34*(5), 301–310.

[24] Chapman, W. W., Bridewell, W., Hanbury, P., Cooper, G. F., & Buchanan, B. G. (2001). Evaluation of negation phrases in narrative clinical reports. In *Proceedings of the AMIA Symposium*. American Medical Informatics Association.

[25] Chen, Q., Allot, A., & Lu, Z. (2020). Keep up with the latest coronavirus research. *Nature*, *579*(7798), 193.

[26] Cheshire, J., Tieken–Boon van Ostade, I., Tottie, G., & Van der Wurff, W. (1999). English negation from an interactional perspective. *Negation in the History of English*, 29–53.

[27] Chiu, K., Grundy, Q., & Bero, L. (2017). 'Spin' in published biomedical literature: A methodological systematic review. *PLoS Biology*, *15*(9), e2002173.

[28] Coates, J. (1987). Epistemic modality and spoken discourse. *Transactions of the Philological Society*, *85*(1), 110–131.

[29] Cofield, S. S., Corona, R. V., & Allison, D. B. (2010). Use of causal language in observational studies of obesity and nutrition. *Obesity Facts*, *3*(6), 353–356.

[30] Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, *20*(1), 37–46.

[31] Cohen, W. W. (1995). Fast effective rule induction In Machine Learning Proceedings of the T welfth International Conference. *Lake Tahoe California Morgan*.

[32] Condoravdi, C., & Lauer, S. (2012). Imperatives: Meaning and illocutionary force. *Empirical Issues in Syntax and Semantics*, *9*, 37–58.

[33] Coyne, J. C., & Tennen, H. (2010). Positive psychology in cancer care: Bad science, exaggerated claims, and unproven medicine. *Annals of Behavioral Medicine*, *39*(1), 16–26.

[34] Crompton, P. (1997). Hedging in academic writing: Some theoretical problems. *English for Specific Purposes*, *16*(4), 271–287.

[35] Cummings, P. (2007). Policy recommendations in the discussion section of a research article. *Injury Prevention*, *13*(1), 4–5.

[36] De Marneffe, M. C., Rafferty, A. N., & Manning, C. D. (2008, June). Finding contradictions in text. In *Proceedings of ACL–08: HLT*, pages 1039–1047, Columbus, Ohio. Association for Computational Linguistics

[37] Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019, January). BERT: In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

[38] Dong, L., Wei, F., Duan, Y., Liu, X., Zhou, M., & Xu, K. (2013, July). The automated acquisition of suggestions from tweets. In *Proceedings of the Twenty–Seventh AAAI Conference on Artificial Intelligence* (pp. 239–245).

[39] Ervin–Tripp, S. (1976). Speech acts and social learning. *Meaning in Anthropology*, 123–153.

[40] Fancellu, F., Lopez, A., & Webber, B. (2016, August). Neural networks for negation scope detection. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (volume 1: long papers)* (pp. 495–504).

[41] Faraoni, D., & Schaefer, S. T. (2016). Randomized controlled trials vs. observational studies: why not just live together?. *BMC Anesthesiology*, *16*(1), 1–4.

[42] Farkas, R., Vincze, V., Móra, G., Csirik, J., & Szarvas, G. (2010, July). The CoNLL–2010 shared task: learning to detect hedges and their scope in natural language text. In: *Proceedings of the Fourteenth Conference on Computational Natural Language Learning — Shared Task* (pp. 1–12). Association for Computational Linguistics.

[43] Faulkner, A. (2014). Automated classification of stance in student essays: An approach using stance target information and the Wikipedia link–based measure. *Science*, *376*(12), 86.

[44] Ferreira, W., & Vlachos, A. (2016, June). Emergent: a novel data–set for stance classification. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics.

[45] Fontelo, P., & Liu, F. (2018). A review of recent publication trends from top publishing countries. *Systematic Reviews*, *7*(1), 1–9.

[46] Fraser, B. (2010). Pragmatic competence: The case of hedging. *New Approaches to Hedging*, *1534*.

[47] Gautret, P., Lagier, J.C., Parola, P., Meddeb, L., Mailhe, M., Doudier, B., Courjon, J., Giordanengo, V., Vieira, V.E., Dupont, H.T. and Honoré, S. (2020). Hydroxychloroquine and azithromycin as a treatment of COVID–19: results of an open–label non–randomized clinical trial. *International Journal of Antimicrobial Agents*, *56*(1), 105949.

[48] Georgescul, M. (2010, July). A hedgehop over a max–margin framework using hedge cues. In *Proceedings of the Fourteenth Conference on Computational Natural Language Learning – Shared Task* (pp. 26– 31). Association for Computational Linguistics.

[49] Glover, J., Izzo, D., Odato, K., & Wang, L. (2006). EBM pyramid and EBM page generator. *New Haven,* CT: Yale University.

[50] Goldin, I., & Chapman, W. W. (2003, July). Learning to detect negation with 'not'in medical texts. In *Proceedings of Workshop on Text Analysis and Search for Bioinformatics, ACM SIGIR*.

[51] Goldberg, A. B., Fillmore, N., Andrzejewski, D., Xu, Z., Gibson, B., & Zhu, X. (2009, June). May all your wishes come true: A study of wishes and how to recognize them. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of*

the Association for Computational Linguistics (pp. 263–271). Association for Computational Linguistics.

[52] Goryachev, S., Sordo, M., Zeng, Q. T., & Ngo, L. (2006). *Implementation and evaluation of four different methods of negation detection* (pp. 2826–2831). Technical report, DSG.

[53] Green, M. C., & Brock, T. C. (2000). The role of transportation in the persuasiveness of public narratives. *Journal of Personality and Social Psychology*, *79*(5), 701.

[54] Halliday, M. A. K. (1970). Language structure and language function. *New Horizons in Linguistics*, *1*, 140– 165.

[55] Halpern, D. F. (1998). Teaching critical thinking for transfer across domains: Disposition, skills, structure training, and metacognitive monitoring. *American Psychologist*, *53*(4), 449.

[56] Haneef, R., Lazarus, C., Ravaud, P., Yavchitz, A., & Boutron, I. (2015). Interpretation of results of studies evaluating an intervention highlighted in Google health news: a cross–sectional study of news. *PLoS One*, *10*(10), e0140889.

[57] Harabagiu, S., Hickl, A., & Lacatusu, F. (2006, July). Negation, contrast and contradiction in text processing. In *AAAI* (Vol. 6, pp. 755–762).

[58] Harkema, H., Dowling, J. N., Thornblade, T., & Chapman, W. W. (2009). ConText: an algorithm for determining negation, experiencer, and temporal status from clinical reports. *Journal of Biomedical Informatics*, *42*(5), 839–851.

[59] Huang, Y., & Lowe, H. J. (2007). A novel hybrid approach to automated negation detection in clinical radiology reports. *Journal of the American Medical Informatics Association*, *14*(3), 304–311.

[60] Heng, C. S. & Tan, H. (2000). Maybe, perhaps, I believe, you could – making claims and the use of hedges. *Second Language Studies*, 19(1). 127–157.

[61] Hope, T., Portenoy, J., Vasan, K., Borchardt, J., Horvitz, E., Weld, D. S., Hearst, M.A., & West, J. (2020). SciSight: Combining faceted navigation and research group detection for COVID–19 exploratory scientific search. *In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (Systems Demonstrations)*, Association for Computational Linguistics.

[62] Hu, G., & Cao, F. (2011). Hedging and boosting in abstracts of applied linguistics articles: A comparative study of English-and Chinese-medium journals. *Journal of Pragmatics*, *43*(11), 2795–2809.

[63] Hyland, K. (1994). Hedging in academic writing and EAF textbooks. *English for Specific Purposes*, *13*(3), 239–256.

[64] Hyland, K. (1995). The Author in the text: Hedging scientific writing. *Linguistics and Language Teaching*, *18*, 33–42.

[65] Hyland, K. (1996). Writing without conviction? Hedging in science research articles. *Applied Linguistics*, *17*(4), 433–454.

[66] Hyland, K., & Milton, J. (1997). Qualification and certainty in L1 and L2 students' writing. *Journal of Second Language Writing*, *6*(2), 183–205.

[67] Hyland, K. (1998a). *Hedging in Scientific Research Articles*. Amsterdam: John Benjamins Publishing.

[68] Hyland, K. (1998b). Boosting, hedging and the negotiation of academic knowledge. *Text–Interdisciplinary Journal for the Study of Discourse*, *18*(3): 349–382.

[69] Ioannidis, J. P. (2005). Why most published research findings are false. *PLoS Medicine*, *2*(8), e124.

[70] Johnson, R. A., & Wichern, D. W. (2014). *Applied multivariate statistical analysis* (Vol. 6). London, UK: Pearson.

[71] Kabisch, M., Ruckes, C., Seibert-Grafe, M., & Blettner, M. (2011). Randomized controlled trials: part 17 of a series on evaluation of scientific publications. *Deutsches Ärzteblatt International*, *108*(39), 663.

[72] Karkkainen, E. (1992). Modality as a strategy in interaction: Epistemic modality in the language of native and non-native Speakers of English. *Pragmatics and Language Learning, 3*, 197–216.

[73] Kilicoglu, H., & Bergler, S. (2008). Recognizing speculative language in biomedical research articles: a linguistically motivated perspective. *BMC Bioinformatics, 9*(11), S10.

[74] Kilicoglu, H., Peng, Z., Tafreshi, S., Tran, T., Rosemblat, G., & Schneider, J. (2019). Confirm or refute?: A comparative study on citation sentiment classification in clinical research publications. *Journal of Biomedical Informatics*, 91, 103123.

[75] Kleinert, S., & Wager, E. (2010, July). Responsible research publication: international standards for editors. In *a Position Statement Developed at the 2nd World Conference on Research Integrity* (pp. 22–24).

[76] Kranich, S. (2009). Epistemic modality in English popular scientific texts and their German translations. *Trans–kom*, *2*(1), 26–41.

[77] Kwong, H., & Yorke–Smith, N. (2012). Detection of imperative and declarative question–answer pairs in email conversations. *AI Communications*, *25*(4), 271–283.

[78] Lakoff, R. (1972) The pragmatics of modality. In: *P. Peranteau, J. Levi, and G. Phares (eds.), Papers from the Eighth Regional Meeting (pp.* 229–46). Chicago Linguistics Society.

[79] Lazarus, C., Haneef, R., Ravaud, P., & Boutron, I. (2015). Classification and prevalence of spin in abstracts of non–randomized studies evaluating an intervention. *BMC Medical Research Methodology, 15*(1), 85.

[80] Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C. H., & Kang, J. (2020). BioBERT: a pre–trained biomedical language representation model for biomedical text mining. *Bioinformatics*, *36*(4), 1234–1240.

[81] Lendvai, P., & Reichel, U. D. (2016). Contradiction Detection for Rumorous Claims. *arXiv preprint arXiv:1611.02588*.

[82] Lewis, K., Chaudhuri, D., Alshamsi, F., Carayannopoulos, L., Dearness, K., Chagla, Z., Alhazzani, W. and GUIDE Group. (2021). The efficacy and safety of hydroxychloroquine for COVID–19 prophylaxis: a systematic review and meta–analysis of randomized trials. *PloS One*, *16*(1), e0244778.

[83] Li, X., Gao, W., & Shavlik, J. W. (2014). Detecting semantic uncertainty by learning hedge cues in sentences using an HMM. In *Workshop on Semantic Matching in Information Retrieval (SMIR) (*p. 11).

[84] Li, Y., Zhang, J., & Yu, B. (2017, September). An NLP analysis of exaggerated claims in science news. In *Proceedings of the 2017 Empirical Methods in Natural Language Processing Workshop: Natural Language Processing Meets Journalism* (pp. 106–111).

[85] Li, Y., Wang, J., & Yu, B. (2021, November). Detecting Health Advice in Medical Research Literature. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing* (pp. 6018–6029).

[86] Li, Y., & Yu, B. (2022, February). Advice Giving in Medical Research Literature. In *International Conference on Information* (pp. 261–272). Springer, Cham.

[87] Light, M., Qiu, X. Y., & Srinivasan, P. (2004, May). The language of bioscience: Facts, speculations, and statements in between. In: *Proceedings of BioLink 2004 Workshop on Linking Biological Literature, Ontologies and Databases: Tools for Users* (pp. 17–24). Association for Computational Linguistics.

[88] Liu, J., Wang, S., & Sun, Y. (2019, June). Olenet at semeval–2019 task 9: Bert based multi–perspective models for suggestion mining. In *Proceedings of the 13th International Workshop on Semantic Evaluation* (pp. 1231–1236).

[89] Lumbreras, B., Parker, L. A., Porta, M., Pollán, M., Ioannidis, J. P., & Hernández-Aguado, I. (2009). Overinterpretation of clinical applicability in molecular diagnostic research. *Clinical Chemistry*, 55(4), 786-794.

[90] Mann, C. J. (2003). Observational research methods. Research design II: cohort, cross sectional, and case–control studies. *Emergency Medicine Journal*, *20*(1), 54–60.

[91] Mao, F., Mercer, R. E., & Xiao, L. (2014, June). Extracting imperatives from wikipedia article for deletion discussions. In *Proceedings of the First Workshop on Argumentation Mining* (pp. 106–107).

[92] Mauranen, A. (2004). "They're a little bit different". Variation in hedging in academic speech. In K. Aijmer & A–B. Stenström (Eds.), *Discourse Patterns in Spoken and Written Corpora*. Amsterdam: John Benjamins.

[93] McDonald, C. J., & Overhage, J. M. (1994). Guidelines you can follow and can trust: an ideal and an example. *JAMA*, *271*(11), 872–873.

[94] McGrath, T. A., Alabousi, M., Skidmore, B., Korevaar, D. A., Bossuyt, P. M., Moher, D., Thombs, B., & McInnes, M. D. (2017). Recommendations for reporting of systematic reviews and meta-analyses of diagnostic test accuracy: a systematic review. *Systematic Reviews*, *6*(1), 194.

[95] Medlock, B., & Briscoe, T. (2007). Weakly supervised learning for hedge classification in scientific literature. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics* (pp. 992–999).

[96] Medlock, B. (2008). Exploring hedge identification in biomedical literature. *Journal of Biomedical Informatics*, *41*(4), 636–654.

[97] Michie, S., & Johnston, M. (2004). Changing clinical behaviour by making guidelines specific. *BMJ*, *328*(7435), 343–345.

[98] Morante, R., & Blanco, E. (2012). SEM 2012 shared task: Resolving the scope and focus of negation. In *SEM 2012: The First Joint Conference on Lexical and Computational Semantics–Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval 2012)* (pp. 265–274).

[99] Morante, R., & Daelemans, W. (2009, June). Learning the scope of hedge cues in biomedical texts. In *Proceedings of the Workshop on Current Trends in Biomedical Natural Language Processing* (pp. 28–36). Association for Computational Linguistics.

[100] Murad, M. H., Asi, N., Alsawas, M., & Alahdab, F. (2016). New evidence pyramid. *BMJ Evidence-Based Medicine*, *21*(4), 125–127.

[101] Mushin, I. (2001). *Evidentiality and epistemological stance: Narrative retelling* (Vol. 87). Amsterdam, Netherlands: John Benjamins Publishing.

[102] Myers, G. (1989). The pragmatics of politeness in scientific articles. *Applied Linguistics*, 10:1–35.

[103] Namsaraev, V. (1997). Hedging in Russian academic writing in sociological texts. *Research in Text Theory*, 64–82.

[104] Neff, J., Dafouz, E., Herrera, H., Martínez, F., Rica, J. P., Díez, M., Prieto, R., & Sancho, C. (2003). Contrasting learner corpora: the use of modal and reporting verbs in the expression of writer stance. *Language and Computers*, *48*, 211–230.

[105] Negi, S. (2016, August). Suggestion mining from opinionated text. In *Proceedings of the ACL 2016 Student Research Workshop* (pp. 119–125).

Negi, S., Daudert, T., & Buitelaar, P. (2019, June). Semeval–2019 task 9: Suggestion mining from online reviews and forums. In *Proceedings of the 13th International Workshop on Semantic Evaluation* (pp. 877–887).

[106] Ochodo, E. A., de Haan, M. C., Reitsma, J. B., Hooft, L., Bossuyt, P. M., & Leeflang, M. M. (2013). Overinterpretation and misreporting of diagnostic accuracy studies: evidence of "spin". *Radiology*, *267*(2), 581–588.

[107] Palmer, F. (1986). *Mood and Modality.* Cambridge: Cambridge University Press

[108] Park, C., Kim, J., Lee, H. G., Amplayo, R. K., Kim, H., Seo, J., & Lee, C. (2019, June). This Is Competition at SemEval-2019 Task 9: BERT is unstable for out-of-domain samples. In *Proceedings of the 13th International Workshop on Semantic Evaluation* (pp. 1254-1261).

[109] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., & Vanderplas, J. (2011). Scikit–learn: Machine learning in Python. *The Journal of Machine Learning Research*, *12*, 2825–2830.

[110] Pillar, C. (2020, April 7th) Former FDA leaders decry emergency authorization of malaria drugs for coronavirus. URL: https://www.sciencemag.org/news/2020/04/former–fda–leaders–decry–emergency–authorization–malaria–drugs–coronavirus, retrieved on April, 20th, 2022.

[111] Pless, I. B. (2009). Three basic convictions: a recipe for preventing child injuries. *Bulletin of the World Health Organization, 87(*5*)*, 395.

[112] Prasad, V., Jorgenson, J., Ioannidis, J. P., & Cifu, A. (2013). Observational studies often make clinical practice recommendations: an empirical evaluation of authors' attitudes. *Journal of Clinical Epidemiology, 66*(4), 361–366.

[113] Qi, P., Zhang, Y., Zhang, Y., Bolton, J., & Manning, C. D. (2020, July). Stanza: A Python Natural Language Processing Toolkit for Many Human Languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations* (pp. 101-108).

[114] Qian, Z., Li, P., Zhu, Q., Zhou, G., Luo, Z., & Luo, W. (2016, November). Speculation and negation scope detection via convolutional neural networks. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing* (pp. 815–825).

[115] Ramanand, J., Bhavsar, K., & Pedanekar, N. (2010, June). Wishful thinking–finding suggestions and'buy'wishes from product reviews. In *Proceedings of the NAACL HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text* (pp. 54–61).

[116] Read, J., Velldal, E., Cavazza, M., & Georg, G. (2016). A corpus of clinical practice guidelines annotated with the importance of recommendations. *In: LREC 2016, Tenth International Conference on Language Resources and Evaluation, 23–28 May 2016, Portorož, Slovenia*

[117] Rizomilioti, V. (2006). Exploring epistemic modality in academic discourse using corpora. *Information Technology in Languages for Specific Purposes* (pp. 53–71). Boston, Massachusetts: Springer.

[118] Reitan, J., Faret, J., Gambäck, B., & Bungum, L. (2015, September). Negation scope detection for twitter sentiment analysis. *In Proceedings of the 6th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis* (pp. 99–108).

[119] Robinson, D. H., Levin, J. R., Thomas, G. D., Pituch, K. A., & Vaughn, S. (2007). The incidence of "causal" statements in teaching–and–learning research journals. *American Educational Research Journal*, 44(2), 400–413.

[120] Sackett, D. L., Rosenberg, W. M., Gray, J. M., Haynes, R. B., & Richardson, W. S. (1996). Evidence based medicine: What it is and what it isn't. *BMJ*, 312(7023), 71–72.

[121] Saha, S., Saint, S., & Christakis, D. A. (2003). Impact factor: a valid measure of journal quality?. *Journal of the Medical Library Association*, 91(1), 42.

[122] Salager–Meyer, F. (1994). Hedges and textual communicative function in medical English written discourse. *English for Specific Purposes*, *13*(2), 149–170.

[123] Schlesselman, J. J. (1982). *Case-Control Studies: Design, Conduct, Analysis*. Oxford University Press.

[124] Schünemann, H. J., Fretheim, A., & Oxman, A. D. (2006). Improving the use of research evidence in guideline development: 1. Guidelines for guidelines. *Health Research Policy and Systems*, *4*(1), 13.

[125] Searle, J. R. (1976). A classification of illocutionary acts. *Language in Society*, 1–23.

[126] Shah, P., Michal, A., Ibrahim, A., Rhodes, R., & Rodriguez, F. (2017). What makes everyday scientific reasoning so challenging?. In *Psychology of Learning and Motivation* (Vol. 66, pp. 251–299). Academic Press.

[127] Shekelle, P. G., Kravitz, R. L., Beart, J., Marger, M., Wang, M., & Lee, M. (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.

[128] Shrestha, L., & McKeown, K. (2004). Detection of question-answer pairs in email conversations. In *COLING 2004: Proceedings of the 20th International Conference on Computational Linguistics* (pp. 889–895).

[129] Somasundaran, S., & Wiebe, J. (2010, June). Recognizing stances in ideological on-line debates. In *Proceedings of the NAACL HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text* (pp. 116–124). Association for Computational Linguistics.

[130] Song, J. W., & Chung, K. C. (2010). Observational studies: cohort and case–control studies. *Plastic and Reconstructive Surgery*, *126*(6), 2234.

[131] Sumner, P., Vivian–Griffiths, S., Boivin, J., Williams, A., Venetis, C. A., Davies, A., Ogden, J., Whelan, L., Hughes, B., Dalton, B., & Boy, F. (2014). The association between exaggeration in health related science news and academic press releases: retrospective observational study. *BMJ*, *349*, g7015.

[132] Szarvas, G., Vincze, V., Farkas, R., & Csirik, J. (2008, June). The BioScope corpus: annotation for negation, uncertainty and their scope in biomedical texts. In *Proceedings of the Workshop on Current Trends in Biomedical Natural Language Processing* (pp. 38–45). Association for Computational Linguistics.

[133] Tang, B., Wang, X., Wang, X., Yuan, B., & Fan, S. (2010, July). A cascade method for detecting hedges and their scope in natural language text. In *Proceedings of the Fourteenth Conference on Computational Natural Language Learning – Shared Task* (pp. 13–17). Association for Computational Linguistics.

[134] Thiese, M. S. (2014). Observational and interventional study design types; an overview. *Biochemia Medica*, *24*(2), 199–210.

[135] Velldal, E. (2011). Predicting speculation: A simple disambiguation approach to hedge detection in biomedical literature. *Journal of Biomedical Semantics*, *2*(5), S7.

[136] Vincze, V., Szarvas, G., Farkas, R., Móra, G., & Csirik, J. (2008). The BioScope corpus: biomedical texts annotated for uncertainty, negation and their scopes. *BMC Bioinformatics*, *9*(11), S9.

[137] Volk, R. J., Jibaja-Weiss, M. L., Hawley, S. T., Kneuper, S., Spann, S. J., Miles, B. J., & Hyman, D. J. (2008). Entertainment education for prostate cancer screening: a randomized trial among primary care patients with low health literacy. *Patient Education and Counseling*, *73*(3), 482–489.

[138] Wei, Z., Chen, J., Gao, W., Li, B., Zhou, L., He, Y., & Wong, K. F. (2013). An empirical study on uncertainty identification in social media context. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)* (Vol. 2, pp. 58–62).

[139] Wharton, S. (2012). Epistemological and interpersonal stance in a data description task: Findings from a discipline–specific learner corpus. *English for Specific Purposes*, *31*(4), 261–270.

[140] Wicaksono, A. F., & Myaeng, S. H. (2012, October). Mining advices from weblogs. In *Proceedings of the 21st ACM International Conference on Information and Knowledge Management* (pp. 2347–2350).

[141] Wicaksono, A. F., & Myaeng, S. H. (2013, June). Automatic extraction of advice-revealing sentences for advice mining from online forums. In *Proceedings of the seventh international conference on Knowledge capture* (pp. 97–104).

[142] Wilson, M. K., & Chestnutt, I. G. (2016). Prevalence of recommendations made within dental research articles using uncontrolled intervention or observational study designs. *Journal of Evidence Based Dental Practice*, *16*(1), 1–6.

[143] Woolf, S., Schünemann, H. J., Eccles, M. P., Grimshaw, J. M., & Shekelle, P. (2012). Developing clinical practice guidelines: types of evidence and outcomes; values and economics, synthesis, grading, and presentation and deriving recommendations. *Implementation Science*, *7*(1), 61.

[144] Xiao, Y., Slaton, Z. Y., & Xiao, L. (2020, May). TV-AfD: An Imperative-Annotated Corpus from The Big Bang Theory and Wikipedia's Articles for Deletion Discussions. In *Proceedings of the 12th Language Resources and Evaluation Conference* (pp. 6542–6548).

[145] Yang, H., De Roeck, A., Gervasi, V., Willis, A., & Nuseibeh, B. (2012, September). Speculative requirements: Automatic detection of uncertainty in natural language requirements. In *Requirements Engineering Conference (RE), 2012 20th IEEE International* (pp. 11–20). IEEE.

[146] Yu, B., Li, Y., & Wang, J. (2019, November). Detecting Causal Language Use in Science Findings. 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)* (pp. 4656–4666).

[147] Yue, P., Wang, J., & Zhang, X. (2019, June). YN-HPCC at SemEval–2019 Task 9: Using a BERT and CNN-BiLSTM-GRU model for suggestion mining. In *Proceedings of the 13th International Workshop on Semantic Evaluation* (pp. 1277–1281).

[148] Zuck, J. G., & Zuck, L. V. (1986). Hedging in newswriting. *Beads or Bracelets*, 172–180.

Author’s Biography

Yingya Li was born in Chengdu, Sichuan province, China. She graduated from the University of Science and Technology Beijing in 2013 with a bachelor’s degree of English Literature. She completed a Master of Linguistic Studies at Syracuse University in 2015. She is currently a postdoc fellow in Harvard Medical School and Boston Children’s Hospital.