Understanding and Mitigating Bias in Online Health Search

Anat Hashavit[1], Hongning Wang[2], Raz Lin[1], Tamar Stern[1], Sarit Kraus[1]
[1]Department of Computer Science Bar Ilan University, Ramat Gan, Israel
[2]Department of Computer Science, University of Virginia, Charlottesville, VA, USA
{anat.hashavit,stern.tamar96}@gmail.com, {linraz,sarit}@cs.biu.ac.il
hw5x@virginia.edu,

ABSTRACT

Search engines are perceived as a reliable source for general information needs. However, finding the answer to medical questions using search engines can be challenging for an ordinary user. Content can be biased and results may present different opinions. In addition, interpreting medically related content can be difficult for users with no medical background. All of these can lead users to incorrect conclusions regarding health related questions. In this work we address this problem from two perspectives. First, to gain insight on users' ability to correctly answer medical questions using search engines, we conduct a comprehensive user study. We show that for questions regarding medical treatment effectiveness, participants struggle to find the correct answer and are prone to overestimating treatment effectiveness. We analyze participants' demographic traits according to age and education level and show that this problem persists in all demographic groups. We then propose a semi-automatic machine learning approach to find the correct answer to queries on medical treatment effectiveness as it is viewed by the medical community. The model relies on the opinions presented in medical papers related to the queries, as well as features representing their impact. We show that, compared to human behaviour, our method is less prone to bias. We compare various configurations of our inference model and a baseline method that determines treatment effectiveness based solely on the opinion of medical papers. The results bolster our confidence that our approach can pave the way to developing automatic bias-free tools that can help mediate complex health related content to users.

CCS CONCEPTS

• **Information systems → Evaluation of retrieval results**; **Clustering and classification**.

KEYWORDS

Health Search, Biases, Machine Learning

ACM Reference Format:
Anat Hashavit[1], Hongning Wang[2], Raz Lin[1], Tamar Stern[1], Sarit Kraus[1], [1]Department of Computer Science Bar Ilan University, Ramat Gan, Israel, [2]Department of Computer Science, University of Virginia, Charlottesville, VA, USA, {anat.hashavit,stern.tamar96}@gmail.com, {linraz,sarit}@cs.biu.ac.il, hw5x@virginia.edu, . 2021. Understanding and Mitigating Bias in Online Health Search . In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '21), July 11–15, 2021, Virtual Event, Canada.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3404835.3462930

SIGIR '21, July 11–15, 2021, Virtual Event, Canada.

ACM ISBN 978-1-4503-8037-9/21/07…$15.00
https://doi.org/10.1145/3404835.3462930

1 INTRODUCTION

According to the most recent PEW research study [2], nearly six in ten Americans query the internet for health related information, out of which 77% start their search process by querying known search engines such as Google, Yahoo or Bing. However, it has been shown that search engines can present results that are biased towards a positive outcome [18]. Bias, as it was defined in [18], is when "*search results describe a deviation from a known or accepted truth that negatively affects result accuracy*". Even if the search results are not biased, understanding medical related content can be complex without the relevant education, especially for content retrieved from information sources aimed at a professional crowd. Such content can appear in commercial search engine results. For example, medical research papers indexed by the *PubMed* corpus and search engine are retrievable by *Google*.

In this work we focus on treatment effectiveness queries. A treatment effectiveness query (TEQ) is a query that tries to answer the question "Is x effective in treating y?", where x is a treatment and y is a medical condition. When search results in TEQs are skewed towards positive outcomes, users might conclude that ineffective treatments are in fact effective, which can be counterproductive to their condition.

The research questions we answer in this paper are:

(1) How susceptible are users to bias in online health search?
(2) Are scientific search engines less prone to bias than general purpose search engines?
(3) Can automatic methods be leveraged to mitigate the bias in online health search?

To answer our research questions we first obtain a list of TEQs and their associated effectiveness labels. We employ medical doctors to determine the effectiveness of each TEQ. These labels are considered to be the ground truth throughout the paper. To answer question Q1 we conduct a study where users are requested to frame and answer TEQs using a search engine of their choice. We analyze and report participants' accuracy, overestimation and underestimation errors. We show that participants are prone to overestimating the effectiveness of treatments. We perform a demographic analysis in which we inspect participants' age group and education level and show that this is not an isolated problem of a single age group or people without higher level education.

We inspect web domains from which participants retrieve their answers and show that even though answers are retrieved from reliable websites, such as *healtline.com* and *PubMed*, participants still tend to overestimate treatment effectiveness. Whether this tendency stems from content bias, participants' bias or a difficulty in thoroughly understanding medically related content, the user study results demonstrate the need for better mitigation of health related information to searchers.

To answer questions Q2 and Q3, we retrieve a set of medical papers from the medical corpus and search engine *PubMed* for each labeled TEQ. We use annotators to extract the stance of the papers for each related TEQ. Based on the set of annotated medical papers we develop the medical treatment effectiveness classifier (MTEC). This is a novel method to semi-automatically assess the effectiveness of medical treatments as viewed by the majority of the medical community. We do so by building a supervised learning-based inference model that tries predict the correct answer for TEQs. The model takes the opinions presented in medical papers relating to the queries, as well as features representing their impact. Single paper features are aggregated to create a rich query feature set and a classifier is used to infer the leading opinion. MTEC does not require the involvement of medical doctors and does not examine the quality of medical research by its content, making it a feasible approach to implement at scale. We compare MTEC to a baseline method that relies only on the opinion of the papers retrieved from *PubMed*. We show that the bias towards positive results presented in previous work is evident in *PubMed* as well, and that it is significantly more likely to overestimate the effectiveness of medical treatments when relying solely on the opinion presented in the retrieved medical papers.

The rest of this paper is organized as follows. Section 2 describes related work in health related IR and content bias. Section 3 presents the user study we conducted. In Section 4 we describe a method to infer treatment effectiveness using supervised learning. Section 5 presents experimental results that evaluate bias in the *PubMed* search engine as well as the performance of various settings of the inference model. Finally, discussions and conclusions are laid out in Section 6.

2 RELATED WORK

As search engines continue to dominate online information retrieval, it is important to understand whether the results are biased, as bias in retrieved search results can lead to incorrect answers. White and Hassan [18] introduce and examine the notion of *content bias*, that is: deviation of the result from the truth. They analyze it in the context of a given medical domain, by observing the logs of the search engine results and comparing the results with crowdsourcing answers. Similar to our study, they utilize the Cochrane reviews as a measure for authoritative ground truth for the queries in question. Their results demonstrate that the search results are biased toward positive outcomes and they associate it with several factors, including skewed content in the engine index and content matching performed by the engine when answering search requests. In another study, White [17] investigates the issue of bias in yes-no question answering based on online search results of medical conditions. In this case, the ground truth is given by the response of two physicians. Similar results are observed with the favoring of positive results (irrespective of the truth). More so, White shows that almost half of the time the search result was incorrect. Pogacar et al. [13] presented users with bias SERP and showed that this bias effects users‘ decision making in health related questions. Our work inspects users‘ decision making process while interacting with genuine search engines and validates not only the bias in health related search but also its adverse effects.

Bias can also be caused by misinformation spread. Wang et al. [16] conducted a systematic literature review on health-related misinformation spreading and found a broad consensus that misinformation is highly prevalent on social media and tends to be more popular than accurate information. Efforts to mitigate misinformation spread have indeed been focused on social media, mainly for domain specific cases. Ghenai and Mejova [7, 8] trained a supervised classifier to predict the spread of misinformation about Zika fever and cancer on Twitter. Kostkova et al. [9] presented a dashboard for tracking anti-vaccination content, again on Twitter. We propose a general medical misinformation detection method that is applicable to a wider variety of conditions.

Determining treatment effectiveness can be viewed as a special case of a question answering task. General medical question answering tasks have also been studied. Some work has focused on improving the retrieval of medical papers. Yates et al. [19] and Lin [11] constructed a citation network from the *PubMed* medical papers corpus in order to improve medical papers' retrieval quality. Both methods test the relevancy of documents, not their correctness or quality and are efficient as tools for medical professionals in their literary review process but not for the layman user.

The challenges in the medical domain led researchers to pursue the development of medical question answering systems. The Human Behaviour Change Project [12], for example, improves behavioral change treatments' effectiveness by using various IR and ML techniques. MedQA [10] and MEANS [3] are medical question answering systems that use IR, NLP and summarization techniques to provide answers to medical questions. MedQA focuses on answering definitional questions and MEANS answers both factual and Boolean questions. MEANS was evaluated against a corpus of previously answered questions, out of which 20 were Boolean questions. The precision ranged from 45% to 60% depending on the level of query relaxation. The authors stated that the system experienced difficulty in answering questions that required a comparison between papers that presented contradictory results. These Q&A systems are summary based and focus mainly on relevant document retrieval and NLP challenges in parsing natural language questions and retrieving informative passages from documents. Yet, there is little to no discussion about questions for which the answer, as derived from medical papers, is inconclusive.

3 USER STUDY

3.1 Determining The Ground Truth

To assess both human and machine ability to correctly answer TEQs, we constructed a dataset of TEQs for which we obtained the correct answers. Most of our TEQs were received from White & Hassan, from the dataset of their paper on content biased in online networks [18]. All of the queries in our dataset correlated to

Cochrane reviews which evaluated their effectiveness. Cochrane [1] is a charity organization whose mission is "*to promote evidence-informed health decision-making by producing high-quality, relevant, accessible systematic reviews and other synthesized research evidence.*" Cochrane only accepts conflict-free funding so their reviewers can be trusted as unbiased. The reviews in Cochrane are structured in the form 'x for y'. For example, 'Melatonin for the treatment of Jet Lag'. A Cochrane review on a given subject is a process in which a team of professionals manually selects and examines a set of random clinical trials and publishes a report with their conclusions. Each report contains an authors' conclusion section that is easily comprehensible given a reasonable understanding of the English language. For example: "*Melatonin is remarkably effective in preventing or reducing jet lag, and occasional short-term use appears to be safe...*". The conclusions relate both to the findings of the various reviewed trials as well as to the quality of the trials' evidence. Therefore, we divided the effectiveness rating of treatments into five possible classes according to the level of their agreement with the claim 'x is effective in treating y', combining the quality of reviewed trials' evidence with the conclusions they present:

(1) Evidence suggests that x is ineffective in treating y;
(2) Preliminary evidence suggests that x is ineffective in treating y;
(3) Evidence is inconclusive or not enough evidence was collected;
(4) Preliminary evidence suggests that x may be effective in treating y;
(5) Evidence suggests that x is effective in treating y.

We employed three physicians to determine the effectiveness of the collected TEQs based on their associated Cochrane reviews. Each review was read by at least two doctors. If the doctors did not agree a third tie breaker review was obtained. If three of the doctors did not agree on a label, the query was removed from the dataset. Doctors did not communicate with each other regarding the reviews.

In addition, in order to be able to supply a concise answer to the query 'Is x effective in treating y?' we aggregate the 1-5 scale to the following 3-class answer:

- no - x is ineffective in treating y;
- maybe - Evidence is inconclusive or not enough evidence was collected;
- yes - x can be effective in treating y.

Similar to [18], we classify a treatment whose ranking is 1 or 2 as ineffective, a ranking of 3 as inconclusive and a ranking of 4 or 5 as effective. The data set will be published and available to future research.

3.2 User Study Setup

To assess users' ability to retrieve the correct answer to a TEQ we conducted a user study. The study's participants were recruited using the Amazon Mechanical Turk (MTurk) website. Since the majority of health related information on the web is in English, participants were restricted to English speaking countries. As a first step, each participant was requested to enter demographic details. The details requested were: Country of residence, age, gender, education level and field of education. In the next stage participants were presented with stories relating to a medical condition and a treatment. Stories had patterns similar to the following form:

> Your friend/friend's relative is suffering from CONDITION. They are considering using TREATMENT to treat it. Your friend relies on you to research the internet and help them decide whether or not this treatment is effective.

Slight adjustments were made per specific queries, for example, if the condition was specific to a certain age group or gender. If the treatment required a doctor's approval, such as in prescription medication or surgical interventions, the phrasing stated 'Your friend is considering requesting their doctors for TREATMENT'.

After reading the story, participants were requested to determine the treatment's effectiveness using a search engine of their choice. The available options were: *yes, no, maybe* and *don't know*. After selecting an answer participants were requested to provide details about their search process. We requested the following details:

(1) The link from which the answer was retrieved.
(2) The portion of the text that convinced the participant to select their answer.
(3) The query used to retrieve the results.
(4) The number of links inspected before reaching the link which contained the answer.
(5) The number of queries entered to the search engine.

Questions 1 and 2 were introduced for quality control. They were randomly inspected by the authors to make sure that the search process was genuine. The last three questions were introduced to gather insights regarding the participants' search process.

Each participant was presented with four or five stories. Ground truth answers were distributed uniformly in each question set. Participants were paid a minimum of $1 for participating. To encourage participants to retrieve the correct answer we added a bonus plan. For each correct answer participants received 1 bonus point, for each incorrect answer a bonus point was deducted. Each bonus point was worth 20 cents. An answer of 'don't know' was not considered as wrong.

3.3 Participants' Performance Analysis

Forty labeled TEQs were used to create the stories presented to participants. In total 117 participants answered 545 questions. 48% of the participants were men and 52% were women. 80% had at least a high school degree or above and 38% of the participants had a Bachelor's degree or above. These numbers are in close proximity to the education attainment distribution in the American public [5]. Participants used common commercial search engines such as *Google*, *Bing*,*Yahoo* and *DuckDuckGo*. Only 5 answers were retrieved using *Google Scholar*. For the vast majority (87%) of questions, participants found a single query enough to reach content that contained the answer to their question. 9% required two queries and only 4% required more than two queries. This demonstrates that participants managed to find relevant content to their queries quite easily. For 44% of the questions, the participants entered a single link which was sufficient in order to find the answer to their questions, 31% of the questions required two links to be examined, 14% required three links and 11% required more than three links,

indicating that while retrieving relevant content was an easy task, finding an actual answer to their question required more effort.

Only five questions were answered with a "don't know". Out of the remaining questions, 47% were answered accurately, 34% were given an answer that overestimated the treatments' effectiveness according to the ground truth and 19% were given an answer that underestimated treatment effectiveness in relation to the ground truth. Table 2 shows the confusion matrix which shows participants' responses to the various question categories. The rows represent the ground truth and the columns the participants' answers to questions. In total, 189 questions were about effective treatments (in relation to the condition in the question), for 157 questions the ground truth stated that the effectiveness of the treatment could not be determined conclusively, and for 194 questions the ground truth stated that the treatment in question was ineffective.

Domain Analysis. Figure 1a presents accuracy and error rates according to the main websites from which participants found answers to their queries. All top websites were reliable and had high quality content. *PubMed* was the most popular domain with 31% of answers retrieved from it. 12% of the questions were answered by information retrieved from the *cochrane.org* website, 7% were answered using data retrieved from *healthline.com*, 6% from *webmd.com* and 6% from the *mayoclinic.org* website. Answers retrieved from *pubmed.org* had the lowest accuracy and the highest chance of overestimating treatment effectiveness. This correlates to positive bias in *PubMed* retrieved results as later presented in Section 5.3. The Mayo Clinic website is the only one for which answers were more likely to underestimate treatment effectiveness than overestimate it, however due to the size of the sample these differences cannot be considered to be statistically significant.

Demographic Analysis. Figures 1b and 1c present answer accuracy and error rates according to different demographics. Figure 1b inspects performance according to education levels and Figure 1c inspects performance according to age groups. There is no statistically significant difference between the various age groups' performance, not in terms of accuracy or in terms of the different error rates. Participants with a B.A. degree or above had a slightly higher accuracy than the rest of the participants, yet the difference was not significant, indicating that even educated people struggle when it comes to medical content.

Effect of Phrasing on Query Generation. In order to determine whether the phrasing of the story effected the queries that participants entered into the search engine, we conducted a small experiment in which stories were phrased in a negative biased language. For example:

> Your friend's grandfather is suffering from dementia. They are considering requesting his doctors to prescribe melatonin, however they heard it's not effective. Your friend relies on you to do research on the internet and help them confirm that it's not effective.

We did not see any difference in participants' performance for negatively framed queries. Inspecting the queries used by participants showed that participants managed to frame unbiased queries even though the story they were presented with was phrased with a negative bias. For example, queries entered to answer the negatively framed story above include: "*melatonin and dementia*", "*is melatonin effective dementia*", "*is melatonin used to treat dementia*" , "*melatonin for dementia patients*".

The findings of the user study show that participants conducted a thorough search. They understood what was requested of them and framed concise queries which led them to relevant content rather quickly. They entered numerous links if needed and visited high quality websites. Still, the results indicate that participants from various age groups and education levels struggled to arrive at the correct response.

To conclude our analysis we bring a few examples of phrases entered by participants, along with the website domains from which they were retrieved, to justify their answer on the effectiveness of melatonin in treating dementia. The ground truth is that melatonin is not effective in treating dementia. However, looking at the phrases supplied by participants, it is easy to see how one can be mistaken in thinking that the treatment is effective.

Text	Source Domain
Consider melatonin. Melatonin might help improve sleep and reduce sundowning in people with dementia.	mayoclinic.org
Melatonin treatment may be effective for the treatment of dementia-related behavior disturbances. Significantly improved outcomes were found from the meta-analysis of psychopathologic behavior and mood scale scores.	cnfbook.org
Its natural quality is part of what makes it so appealing; however, it may not be appropriate for everyone. Some elderly people, including those with underlying dementia, may increase their risk for serious medical complications by taking melatonin without supervision.	healthfully.com

Table 1: Text entered by users to justify their answer

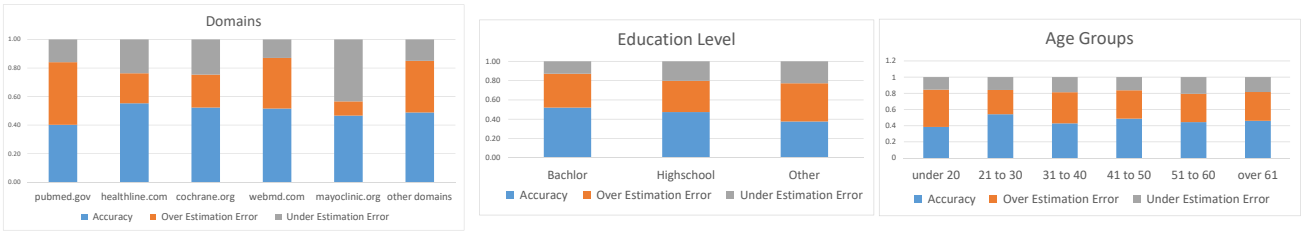
4 LEVERAGING MACHINE LEARNING TO PREDICT TREATMENT EFFECTIVENESS

Given the results in the previous section it is evident that search engine users are susceptible to bias when trying to answer TEQs using said search engines. In this section we show how to mitigate said bias by using machine learning techniques. We present the Medical Treatment Effectiveness Classifier (MTEC). MTEC is a supervised learning-based method that uses features of medical papers to infer a treatment's effectiveness for a given condition.

Our goal is to generate a three-class answer to a TEQ (yes,no,maybe). However, since the scale which describes effectiveness is wider, we

	no	maybe	yes
Ineffective	68	70	56
Inconclusive	27	70	60
Effective	18	55	116

Table 2: Confusion matrix for study participants responses.



(a) Domains **(b) Education Level** **(c) Age Groups**

Figure 1: Answers accuracy and error rate for different demographics and domains.

first describe a classifier that learns an effectiveness ranking classification according to a 1 to K scale. The effectiveness ranking classifier is described in Section 4.1 In Section 4.2 we describe the implementation aspects of MTEC, including the various options of transferring the 1 to K effectiveness ranking to one of the three possible answer classes.

4.1 Effectiveness Ranking Classifier

Given a treatment x and a condition y we define a treatment-condition query $q_{<x,y>}$: 'how effective is x in treating y'. In addition we define the effectiveness ranking of $q_{<x,y>}$, $ER(q_{<x,y>}) \in \{1, .., K\}$. Let $P_{q_{<x,y>}} = \{p_1, .., p_n\}$ be a set of medical papers that discuss $q_{<x,y>}$. The learning task is to infer $ER(q_{<x,y>}|P_{q_{<x,y>}})$.

Each paper $p_i \in P_{q_{<x,y>}}$ has its own stance regarding the effectiveness ranking of $q_{<x,y>}$. Let $\sigma_1, .., \sigma_n$ be the set of stance observations corresponding to $P_{q_{<x,y>}}$ s.t $\sigma_i \in \{1, .., K\}$ represents the stance of paper p_i. The simplest approach would be to use majority voting and select the rating supported by most papers. However, even if $P_{q_{<x,y>}}$ contains all papers relating to $q_{<x,y>}$, which eliminates retrieval bias problems, using a majority vote method is still problematic since it assumes all papers are of equal quality and impact. Yet, the quality of research varies. In order to make a sound judgment one needs to observe not only the number of papers supporting each opinion but also the quality of these papers and the impact they had on the scientific community. To this end we created the effectiveness ranking classifier. Using the set of labeled queries we build a classifier which infers $ER(q_{<x,y>}|\phi_1, .., \phi_K, \phi_q)$ where ϕ_i is the feature set meant to capture the impact of papers supporting stance i and ϕ_q is an additional set of meta features.

The feature generation process consists of three phases. In the first phase, paper level features are collected for each paper. All but one of the features are automatically collected. The only feature requiring human intervention is the stance score feature. The stance score feature represents the opinion of authors regarding the effectiveness of x in treating y. It is annotated by non-expert human annotators that are provided with the abstract of the paper alone. Annotators are not required to be medical experts since their task is merely to understand the conclusions, as described by the authors, and not to examine the quality of the evidence or the methodology that led to these conclusions. The use of non-professional annotators on health related data has been previously attempted in other tasks and has been shown to be effective [15, 20, 21]. The automatically collected features are derived from the paper's publication date, its citation count and the h-index of the journal in which the paper was published. The journal h-index is a citation-based metric that evaluates the scientific impact of a journal [4]. We chose to use this metric since it is widely acceptable and available. In total each paper was described by six different features. Its stance regarding the related query, how many years ago it was published, its h-index, its citation count and two generated features that combined paper's recency with the impact measurements: recency weighted h-index and recency weighted citation count.

Once paper level features are generated, stance level features are constructed. Values of paper level features are grouped according to their associated paper's stance score and added to the query feature vector. In order to better capture the relationship between different stance features we use both the total and mean value of stance features. Finally, query level features are added. The query level features that we chose to add were the total number of relevant papers retrieved as well as the mean value of all paper level features (mean citation count of all papers, mean h-index, etc.).

Once features are constructed a supervised learning classifier is used to generate the inference model. The entire process of generating the inference model is graphically described in Figure 2. It begins by collecting a set of queries for which the treatment's effectiveness rating has been confirmed by a credible source. For each query a collection of medical papers relating to it is retrieved from an online available medical corpus. Features are collected for each paper and aggregated to create a query feature set. A classifier is then trained to learn treatment effectiveness rating.

4.2 Implementing MTEC

4.2.1 Retrieving Relevant Papers. As stated in Section 4.1, a query's feature set is derived from a set of medical papers relating to it. To that end, we implemented a python program that automatically retrieved medical papers from the medical papers corpus and search engine *PubMed*. We chose *PubMed* since it is a widely accepted medical papers corpus and search engine, it has a convenient API and it does not limit or deny automatic crawling. In order to assist the search engine in producing better results we conducted a preliminary pre-processing phase where uninformative phrases were removed from the search query. This included general stop words, as well as words, such as 'treatment' and 'prevents', that are not usually considered to be search engine stop words.

PubMed offers two sorting options. "Most Recent" and "Best Match". We chose to use "Best Match" since the "Most Recent" option returned many irrelevant documents. We restricted the search

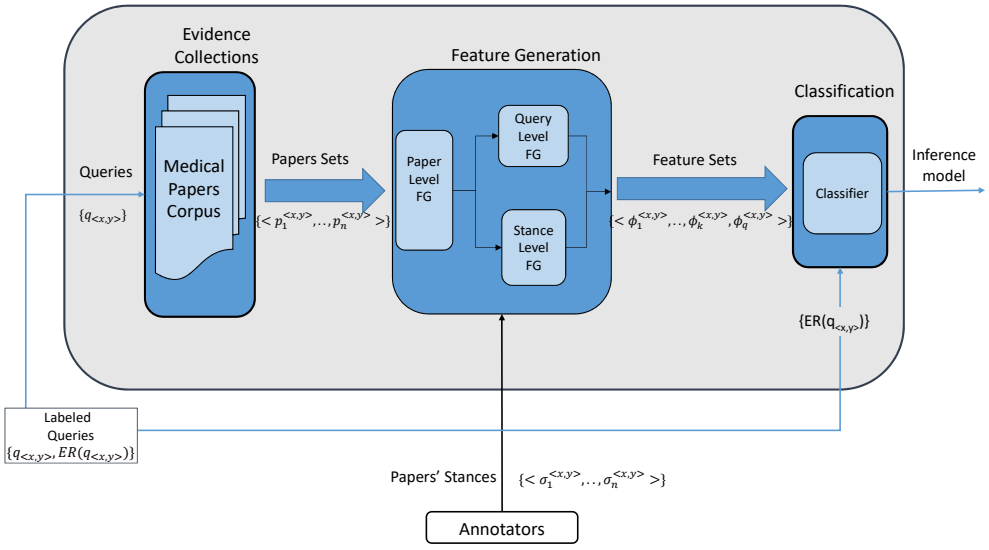


Figure 2: Inference Model Generation - The model receives as input a set of labeled treatment-condition queries and outputs an effectiveness ranking model that predicts effectiveness ranking of previously unseen queries

results to papers published up to 15 years prior to the Cochrane review date of the respective query since it coincides with the average time-span of included papers used by many Cochrane reviews.

We also wanted to maximize the number of relevant papers retrieved while maintaining the overall number of papers to annotate within the limit of our paid annotation budget. We therefore limited the number of papers per query to a maximum of 20 papers. *PubMed* provides a filtering option according to the type of papers. After running some preliminary experiments we chose to retrieve the first 10 clinical trials and the first 10 review papers, again in order to maximize the number of relevant papers retrieved. If the combined number of papers did not reach 20 papers we allowed for an additional unfiltered search.

4.2.2 Paper Stance Score Annotation. To annotate the stance score of papers we used human annotators. All annotators were given a preliminary test of 20 papers relating to a single treatment condition pair. The test results were compared to a previously classified file annotated by the authors. Only candidates with a mean absolute error smaller than 1 were hired. Annotators were not familiar with each other. A single paper had a minimum of three and a maximum of five annotators. In total 26 annotators were used to annotate papers. Their ages spanned from 23 to 60. All of them had an academic education. Six had a PhD and six had a Masters degree. All of the annotators had life science related education, such as a Biology or a Life Sciences degree. None of the classifiers were medical doctors or even medical students.

We chose to use these annotators instead of a crowd-sourcing platform since it was more cost effective in light of the annotators' education level. This conclusion was the result of a preliminary experiment we conducted using Amazon Mechanical Turk.

Annotators were given sets of files in folders whose title names were claims of the form 'x for y' where x was a treatment and y a condition. Each file contained a list of paper abstract URLs from the *PubMed* website. Annotators were requested to rank on a scale of 1 to 5 the level of which the paper's conclusion supports the claim: 'x is effective in treating y'. If annotators could not find a conclusion in the abstract they assigned the paper a negative value. Annotators were instructed to address the conclusion of the authors only; annotators did not have access to the entire paper. Disagreements were settled by majority vote and irrelevant papers were removed from the dataset.

4.2.3 Stance Aggregation and Answer Generation. Our final goal is to generate a concise three-class answer according to the process described in Section 3.1. Since both physician labeling of ground truth and annotators' papers' stance scores are given on a 1-5 scale, a transformation from a 1-5 scale to an answer class needed to be made. There are three ways to do so using the effectiveness ranking classifier: *post-learning transfer*, *label pre-learning transfer* and *label and stance pre-learning transfer*. *post-learning transfer* features sets are aggregated according to a 1-5 scale; labels are also set on a 1-5 scale and thus an effectiveness ranking of 1-5 is learned by the classifier. The answer is then transferred from the inferred effectiveness rating to one of three possible classes. With the *label pre-learning transfer* method the ground truth labels are transferred to answer classes prior to the inference process, thus a three-class answer is directly learned, however the feature sets are still aggregated according to a 1-5 scale. With *label and stance pre-learning transfer*, ground truth labels are transferred to answer classes prior to the inference process. In addition, the stance of papers is also transferred to a 1-3 scale prior to the feature generation process and thus features are aggregated on a 1-3 class according to the possible answer classes. We experimented with all of these settings. Results were not substantially different, however the *label and stance pre-learning transfer* method managed to produce the best results for various forms of classifiers, indicating that the differences between

the two ratings, (1,2) and (4,5), are minor, thus separating them both into the labels and feature sets weakens the generated classifiers.

5 EVALUATING SCIENTIFIC PAPER-BASED METHODS

We now assess the performance of methods that rely on scientific papers. We evaluate both the performance of our machine learning-based method, MTEC, and a baseline method that relies only on the opinion of the papers in *PubMed*, thus evaluating the bias of this scientific search engine.

5.1 Dataset

Our dataset contains 262 TEQ queries, the vast majority of which were obtained courtesy of White & Hassan from the dataset of their paper on content bias in online networks [18]. A smaller number of queries were randomly retrieved from the Cochrane review website by the authors prior to contacting White & Hassan. The queries' effectiveness rating and label were set according to the process describes in section 3.1

We removed queries for which the label was not agreed upon by at least two physicians (13 queries), queries for which none of the retrieved papers specifically discussed the treatments' effectiveness (15 queries), resulting with the remaining 234 queries, out of which 38 were labeled as ineffective, 84 as neutral and 112 as effective.

We removed outliers from the dataset. A query was considered to be an outlier if there was not even a single paper whose stance was in the same label class as the query. Twenty-four queries were identified as outliers. The dataset without outliers contained 30 queries that were labeled as ineffective, 70 inconclusive and 110 effective. Since the dataset was unbalanced we undersampled the effective treatments class so that it had the same number of queries as the inconclusive class. We did not undersample both classes to the size of the ineffective class since it would have resulted in a very small dataset.

5.2 Evaluation Measurements

The confusion matrices for each of the models can be seen in Table 3. Table 4 details performance statistics for each of the classifiers. For each row the best result for that row's statistic is noted in bold. We present for each classifier its general accuracy and per class precision, recall and F-Score metrics [14, p. 5]. Precision estimates the portion of correctly classified examples in a class out of all *predicted* examples in that class, recall estimates the portion of correctly predicted examples in a class out of all examples in that class. F-Score is an harmonic mean between the two, $F\text{-}Score = 2 \cdot \frac{precision \cdot recall}{precision + recall}$.

While in most classification tasks the error of a false-positive prediction is not considered different than the error of a false negative prediction, in a health related search this not the case. In prevention medicine, such as vaccines for example, underestimating the effectiveness of a treatment can cause people to neglect getting vaccinated for preventable diseases, which can be more harmful than, for example, overestimating the effectiveness of some vaccine (notice we are referring to effectiveness, not safety). As our setting discusses treatment effectiveness, predicting an ineffective treatment as effective is more costly since it may cause users to spend money on ineffective treatments (for example food supplements that can be purchased without prescription), or even neglect referring to a doctor for their condition, which may worsen it.

Since our classification task is cost sensitive we introduce two additional metrics for each class: overestimation rate and underestimation rate. The overestimation rate represents the portion of examples in a class that were classified as being more effective than their actual effectiveness class. Accordingly, the underestimation rate represents the portion of examples in a class that were classified as being less effective than their actual effectiveness class. The overestimation rate of the effective class will therefore be 0 for all classifiers and the underestimation rate of the ineffective class will be 0 as well for all classifiers.

5.3 Evaluating Bias in Scientific Search Engines

To answer our research question regarding bias in scientific search engines, we created a simple majority-based classifier based on the papers retrieved from *PubMed*. This is also the baseline method in our analysis of MTEC's performance. The majority classifier simply counts the number of papers supporting each class and then returns the class with the highest number of votes. For example, if a query had 3 relevant papers whose stance scores were 1,2 and 4, the classification of that query, according to majority vote, would be set to the ineffective class since it had 2 votes, in comparison to 1 vote, for the effective class.

As can be observed in Table 3a the majority classifier classified 120 of the 170 queries as positive. In total 35% of the queries in the dataset were overestimated by the classifier. Out of the 30 ineffective treatments in the dataset only 43% were labeled as such, and out of 70 treatments whose effectiveness could not be determined conclusively, 60% were labeled as effective, only 30% were classified accurately, and just 10% were underestimated and classified as ineffective. This means that when assessing a treatment's effectiveness based on the opinion of medical papers alone, unless the treatment is in fact effective, it is extremely likely to arrive at a conclusion biased towards a positive result.

When looking into the distribution of annotated papers, out of 1,778 annotated papers for the 170 queries in the dataset, 267 (15%) papers had a negative opinion regarding the presented treatment effectiveness (stance score of 1,2), 303 (17%) thought that the treatment's effectiveness could not be determined conclusively (stance score of 3) and 1,208 (68%) papers stated that the treatment could be effective (stance score of 4-5). These results again show that positive bias in health search exists in professional search engines as well.

5.4 MTEC Classification Method

To compensate for the small number of ineffective treatments we used Synthetic Minority Oversampling Technique (SMOTE) [6]. SMOTE is a method to generate synthetic examples of the minority class based on features similarity. We found that SMOTE slightly reduced the accuracy for the effective and inconclusive class prediction but improved the overall accuracy by an average of 2%, as well as the recall of ineffective treatments.

Since our dataset was not very large we could not use methods such as deep learning. We attempted various forms of supervised

	Predicted Ineffective	Predicted Inconclusive	Predicted Positive
Ineffective	13	3	14
Inconclusive	7	21	42
Effective	4	2	64

(a) Majority Vote

	Predicted Ineffective	Predicted Inconclusive	Predicted Positive
Ineffective	15	11	4
Inconclusive	7	43	20
Effective	4	8	58

(b) Optimistic Ensemble

	Predicted Ineffective	Predicted Inconclusive	Predicted Positive
Ineffective	20	6	4
Inconclusive	10	48	12
Effective	5	11	54

(c) Random Forest

	Predicted Ineffective	Predicted Inconclusive	Predicted Positive
Ineffective	26	3	1
Inconclusive	19	42	9
Effective	10	14	46

(d) Pessimistic Ensemble

	Predicted Ineffective	Predicted Inconclusive	Predicted Positive
Ineffective	22	7	1
Inconclusive	18	35	17
Effective	7	12	51

(e) KNN

	Predicted Ineffective	Predicted Inconclusive	Predicted Positive
Ineffective	23	6	1
Inconclusive	16	38	16
Positive	4	16	50

(f) Impartial Ensemble

Table 3: Confusion Matrices of the various classifiers for a dataset of 30 ineffective treatments, 70 inconclusive ones and 70 effective treatments.

learning classifiers suited for a dataset of our size. The methods we tried were: random forest classifier, K-Nearest-Neighbor, logistic regression and multi-class SVM. Random forest classifier and KNN (using 5 neighbors) presented the best results.

During our experiments we noticed that different classifiers had different performances for the various classes. The majority-based classifier, for example, had a high precision for the ineffective and inconclusive class, but a very low precision for the effective class and a very low recall for the ineffective class. The KNN classifier, on the other hand, had a high recall for the ineffective class but low precision. Therefore, in light of the cost sensitive nature of our task, we constructed three ensemble classifiers which took into consideration the predictions of all three classifiers according to the cost of prediction errors between the various classes. The three ensemble configurations we used were:

- Pessimistic Ensemble: Considers overestimating a treatment's effectiveness as more costly than underestimating it.
- Optimistic Ensemble: Considers underestimating a treatment's effectiveness as more costly than overestimating it.
- Impartial Ensemble: Considers all prediction errors to be equal.

The pessimistic ensemble considers overestimating a treatment's effectiveness as more costly than underestimating it. Therefore, having to choose between different predictions returned by the three classifiers, it will always prefer the inconclusive classification over the effective classification, and the ineffective classification over the inconclusive classification. Classifying a treatment as effective will require all three classifiers to agree on the classification. Since the majority baseline is extremely overoptimistic, its opinion in the optimistic and impartial ensemble was considered only if it did not classify the treatments as effective. The optimistic ensemble considers underestimating a treatment's effectiveness as more costly than overestimating it. Therefore, having to choose between different predictions returned by the classifiers under consideration, it will always prefer the effective classification over the inconclusive classification, and the inconclusive classification over the ineffective classification. Classifying a treatment as ineffective will require all three classifiers to agree on the classification. The impartial ensemble considers all prediction errors to be equal, therefore, having to choose between different predictions returned by the classifiers under consideration, it will select the prediction with the most classifications. In case of a tie a prediction will be selected between the possibilities uniformly at random.

5.5 Result Analysis and Comparison

All methods achieved a higher accuracy than the user study participants who achieved a general accuracy score of 0.47, an overestimation rate of 0.34 and an underestimation rate of 0.19. We ran a *t-test* comparison between the performance of the user study participants and each of the scientific paper-based methods (including majority classifier) and found the accuracy difference between the methods to be statistically significant ($p\text{-}value < 0.05$).

The majority classifier presented the highest recall for the effective class (0.91) but had a very low precision of 0.53. It also had the highest overestimation rate of 0.35, almost 3 times as much as the random forest classifier's rate of 0.13 and more than 4 times the 0.08 rate of the pessimistic ensemble. It also had relatively low F-Scores compared to the machine learning-based methods.

The best overall accuracy was achieved by the random forest classifier. The random forest classifier also showed the best performance in terms of F-Score for all classes except a very small difference of 0.01 in the ineffective class F-Score in comparison to the impartial ensemble classifier. *T-test* results between random forest accuracy and majority classifier accuracy showed that this difference is statically significant as well. The KNN classifier had a higher recall than the random forest classifier for the ineffective

	Majority Vote	Random Forest	KNN	Pessimistic Ensemble	Optimistic Ensemble	Impartial Ensemble
Accuracy	0.58	**0.72**	0.64	0.67	0.68	0.65
Classifier Overestimation Rate	0.35	0.13	0.15	**0.08**	0.21	0.14
Classifier Underestimation Rate	**0.08**	0.15	0.22	0.25	0.11	0.21
Ineffective Precision	0.54	0.57	0.47	0.47	**0.58**	0.53
Ineffective Recall	0.43	0.67	0.73	**0.87**	0.50	0.77
Ineffective F-Score	0.48	0.62	0.57	0.61	0.54	**0.63**
Ineffective Overestimation Rate	0.57	0.33	0.27	**0.13**	0.50	0.23
Ineffective Underestimation Rate	0	0	0	0	0	0
Inconclusive Precision	**0.81**	0.74	0.65	0.71	0.69	0.63
Inconclusive Recall	0.30	**0.69**	0.50	0.60	0.61	0.54
Inconclusive F-Score	0.44	**0.71**	0.56	0.65	0.65	0.58
Inconclusive Overestimation Rate	0.60	0.17	0.24	**0.13**	0.29	0.23
Inconclusive Underestimation Rate	**0.10**	0.14	0.26	0.27	**0.10**	0.23
Effective Precision	0.53	0.77	0.74	**0.82**	0.71	0.75
Effective Recall	**0.91**	0.77	0.73	0.66	0.83	0.71
Effective F-Score	0.67	**0.77**	0.73	0.73	0.76	0.73
Effective Overestimation Rate	0	0	0	0	0	0
Effective Underestimation Rate	**0.09**	0.23	0.27	0.34	0.17	0.29

Table 4: Classifiers Evaluation Metrics.

class (0.73 vs. 0.67) but a higher inclination to overestimate treatments in the inconclusive class, with an overestimation rate of 0.24 versus only 0.17 of the random forest classifier.

All ensemble methods achieved a higher accuracy in comparison to the majority and KNN classifiers but not in comparison to the random forest classifier. The impartial ensemble approach did not achieve any notable results. The pessimistic ensemble achieved the best recall for the ineffective class (0.87) and the highest prediction for the effective class (0.82). It also had the lowest overestimation rate for all classes. The optimistic ensemble had the second lowest underestimation rates (after the majority classifiers) for all classes but it presented a higher F-Score than the majority classifier for all other classes, showing an inclination towards positive results, while maintaining reasonable performance for the other classification tasks, thus being a better alternative than the majority classifier even when the underestimation errors are more costly.

To summarize we go back to our initial research questions. The performance of the majority baseline classifier as well as our analysis of *PubMed* shows that the answer to Q2 is that, unfortunately, bias is evident in scientific search engines as well. On a more positive note, our results indicated that the answer to Q3 is yes - automatic methods be can leveraged to mitigate the bias in online health search.

6 CONCLUSION

In this paper we explored the phenomenon of bias in online health search. We conducted a user study and showed that although participants conducted a thorough search and visited reliable websites, they still struggled to find correct answers to health related queries. When questioned regarding medical treatments' effectiveness, they were especially prone to overestimating a treatment's effectiveness. We showed that this tendency is not exclusive to a certain age group in the population or to people without a higher level of education. While much work has been done on online bias and misinformation in the health domain, our work shows the impact of this bias on users' decision-making process and emphasises the need for better tools to mediate health related content to users.

To tackle the problem of estimating efficiency of medical treatments we also presented a novel machine learning model which is based on features of published medical papers. Comparing our methods to a baseline approach that merely counts the opinions of the relevant medical papers shows that the latter is prone to a lower accuracy and a higher tendency to overestimate treatment effectiveness. Our results highlight the fact that positive bias is not exclusive to commercial search engines, but exists also in professional content search engines such as *PubMed*. Despite the positive bias in the papers, our machine learning method managed to achieve high accuracy since it considered features such as published journals' h-index and citation count. Moreover, it had a higher accuracy and a smaller overestimation rate in comparison to humans. A key advantage of our method is that, unlike existing Q&A methods in the health domain, our method is not a summarization. It supplies a concise answer to users and does not require the users to have medical understanding or to interpret medically related content. While the ground truth in our dataset was determined by doctors, the stance of papers was annotated by users with no medical degree, showing the promise embodied in our novel method, allowing its implementation on higher scales without employing medical professionals, which are an expensive resource.

7 ACKNOWLEDGMENTS

We thank Ryen White and Ahmed Hassan for sharing their data with us. This work was supported in part by the NSF IIS (Grant No. 1553568) and the Israel Innovation Authority (Grant No. 70069).

REFERENCES

[1] [n.d.]. Cochrane | Trusted evidence. Informed decisions. Better health. https://www.cochrane.org/. (Accessed on 02/06/2021).

[2] 2019. Majority of Adults Look Online for Health Information | PEW Research Center. https://www.pewresearch.org/fact-tank/2013/02/01/majority-of-adults-look-online-for-health-information/. https://www.pewresearch.org/fact-tank/2013/02/01/majority-of-adults-look-online-for-health-information/

[3] Asma Ben Abacha and Pierre Zweigenbaum. 2015. MEANS: A medical question-answering system combining NLP techniques and semantic Web technologies. *Information processing & management* 51, 5 (2015), 570–594.

[4] Lutz Bornmann and Hans-Dieter Daniel. 2007. What do we know about the h index? *Journal of the American Society for Information Science and technology* 58, 9 (2007), 1381–1385.

[5] US Census Bureau. 2020. Educational Attainment in the United States: 2018. https://www.census.gov/data/tables/2018/demo/education-attainment/cps-detailed-tables.html

[6] Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. 2002. SMOTE: synthetic minority over-sampling technique. *Journal of artificial intelligence research* 16 (2002), 321–357.

[7] Amira Ghenai and Yelena Mejova. 2017. Catching Zika Fever: Application of Crowdsourcing and Machine Learning for Tracking Health Misinformation on Twitter. (jul 2017). arXiv:1707.03778 http://arxiv.org/abs/1707.03778

[8] Amira Ghenai and Yelena Mejova. 2018. Fake Cures: User-centric Modeling of Health Misinformation in Social Media. 2, November (2018). arXiv:1809.00557 http://arxiv.org/abs/1809.00557

[9] Patty Kostkova, Vino Mano, Heidi J. Larson, and William S. Schulz. 2016. VAC Medi+board. In *Proceedings of the 6th International Conference on Digital Health Conference - DH '16*. ACM Press, New York, New York, USA, 163–164. https://doi.org/10.1145/2896338.2896370

[10] Minsuk Lee, James Cimino, Hai Ran Zhu, Carl Sable, Vijay Shanker, John Ely, and Hong Yu. 2006. Beyond information retrieval—medical question answering. In *AMIA annual symposium proceedings*, Vol. 2006. American Medical Informatics Association, 469.

[11] Jimmy Lin. 2008. PageRank without hyperlinks: Reranking with PubMed related article networks for biomedical text retrieval. *BMC bioinformatics* 9, 1 (2008), 270.

[12] Susan Michie, James Thomas, Johnston, et al. 2017. The Human Behaviour-Change Project: Harnessing the power of artificial intelligence and machine learning for evidence synthesis and interpretation. *Implementation Science* 12 (12 2017). https://doi.org/10.1186/s13012-017-0641-5

[13] Frances A Pogacar, Amira Ghenai, Mark D Smucker, and Charles LA Clarke. 2017. The Positive and Negative Influence of Search Results on People's Decisions about the Efficacy of Medical Treatments. In *Proceedings of the ACM SIGIR International Conference on Theory of Information Retrieval*. ACM, 209–216.

[14] Hinrich Schütze, Christopher D Manning, and Prabhakar Raghavan. 2007. *An introduction to information retrieval*. Cambridge University Press,.

[15] Joseph D Tucker, Suzanne Day, Weiming Tang, and Barry Bayus. 2019. Crowdsourcing in medical research: concepts and applications. *PeerJ* 7 (2019), e6762.

[16] Yuxi Wang, Martin McKee, Aleksandra Torbica, and David Stuckler. 2019. Systematic literature review on the spread of health-related misinformation on social media. *Social Science & Medicine* 240 (2019), 112552.

[17] Ryen White. 2013. Beliefs and biases in web search. In *Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval*. 3–12.

[18] Ryen W White and Ahmed Hassan. 2014. Content bias in online health search. *ACM Transactions on the Web (TWEB)* 8, 4 (2014), 1–33.

[19] Elliot J Yates and Louise C Dixon. 2015. PageRank as a method to rank biomedical literature by importance. *Source code for biology and medicine* 10, 1 (2015), 16.

[20] Bei Yu, Matt Willis, Peiyuan Sun, and Jun Wang. 2013. Crowdsourcing participatory evaluation of medical pictograms using Amazon Mechanical Turk. *Journal of medical Internet research* 15, 6 (2013), e108.

[21] Haijun Zhai, Todd Lingren, Louise Deleger, Qi Li, Megan Kaiser, Laura Stoutenborough, and Imre Solti. 2013. Web 2.0-based crowdsourcing for high-quality gold standard development in clinical natural language processing. *Journal of medical Internet research* 15, 4 (2013), e73.