Aggregating Personal Health Messages for Scalable Comparative Effectiveness Research

Jason H.D. Cho[★◇] hcho33@illinois.edu
Vera Q.Z. Liao[★◇] liao28@illinois.edu
Yunliang Jiang[‡] jiangyunliang@gmail.com
Bruce R. Schatz[†★◇] schatz@illinois.edu

[★] Dept. of Computer Science, [†]Dept. of Medical Information Science
[◇] University of Illinois at Urbana-Champaign, Urbana, IL, 61801
[‡] Twitter Inc, San Francisco, CA, 94103

ABSTRACT

Comparative Effectiveness Research (CER) is defined as the generation and synthesis of evidence that compares the benefits and harms of different prevention and treatment methods. This is becoming an important field in informing health care providers about the best treatment for individual patients. Currently, the two major approaches in conducting CER are observational studies and randomized clinical trials. These approaches, however, often suffer from either scalability or cost issues.

In this paper, we propose a third approach of conducting CER by utilizing online personal health messages, e.g., comments on online medical forums. The approach is effective in resolving the scalability and cost issues, enabling rapid deployment of system to identify treatments of interests, and developing hypotheses for formal CER studies. Moreover, by utilizing the demographic information of the patients, this approach may provide valuable results on the preferences of different demographic groups. Demographic information is extracted using our high precision automated demographic extraction algorithm. This approach is capable of extracting more than 30% of users' age and gender information.

We conducted CER by utilizing personal health messages on breast cancer and heart disease. We were able to generate statiatically valid results, many of which have already been validated by clinical trials. Others could become hypothesis to be tested in future CER research.

General Terms

CER, Comparative Effectiveness Research, Personal Health Messages, Sentiment Analysis, Demographics Extraction

1. INTRODUCTION

Comparative Effectiveness Research (CER) is defined as the generation and synthesis of evidence that compares the benefits and harms of alternative methods to prevent, diagnose, treat, and monitor clinical conditions or to improve the delivery of care. The ultimate goal of CER is in assisting consumers, clinicians, purchasers and policy makers to make informed decisions that will improve health care at both the individual and population levels [34]. The American Recovery and Reinvestment Act of 2009 (ARRA) allotted $1.1 billion to support this form of research [38]. Various studies have been done in this area, mainly in the form of clinical trials [26] or observational studies [12].

The two approaches to conducting CER, however, are not without their weaknesses. Randomized trials often allow for accurate comparison, but are expensive to conduct [27] and not very scalable, therefore have to be conducted with caution (e.g., with valid, well-motivated hypotheses). Observational studies using patient records, on the other hand, have potential privacy issues that need to be taken into account. Also they are often conducted in distributed research networks (DRNs) and are subject to different state laws and regulations as well as institution-specific policies [23, 40].

We propose using online personal health messages (e.g., online forum posts, blog posts, product reviews) in conducting CER research. These are defined as online posts that are generated by patients and exclude those that are written by experts. The benefits of using personal health messages are four-fold. First, in contrast to the limited granulairty of randomized trials, it is possible to utilize a 'big data' approach to conduct these studies. In each of the population studies that we have conducted, we were able to collect health messages from tens of thousands of patients over a period of five years. On the other hand, experiments are conducted on the order of tens to hundreds in randomized clinical trials, often spanning a shorter period of time. It is further worth noting that Medhelp is not the only source of personal health messages. There are other medical forums such as those from WebMD or HealthBoards which can be aggregated to broaden the cohort pool.

Second, personal health messages may reflect the latest trends in treatments. Studies showed it is possible to cluster or predict adverse drug effects by using health messages [10, 22], many of which were out on the market for only a few years.

Third, personal health messages are publicly available on various sites with various samples, which allows repeated experiments. Neither the results from randomized trials nor

BCB'13 September 22 - 25 2013, Wshington, DC, USA

http://dx.doi.org/10.1145/2506583.2512363.

observational studies can easily be replicated without devoting significant amount of resources.

Finally, online health messages often provide individual contexts such as the demographic information of the message source, either explicitly or implicitly, in the statements or personal profile, which can be extracted by text processing methods. It allows the generation of results on the preferences of different demographic groups, which may solve difficulties faced by traditional CER research. In clinical trials, without ways of generating valid hypotheses, one has to exhaustively examine all groups, which is impractical without a large and well-planned sample.

In summary, CER using personal health messages can be used as a breadth-first-search approach in searching for potential differences in effectiveness. If the results indicate statistical significance, a further in-depth, comprehensive study such as observational, or random trial studies can be conducted, guiding CER research to efficiently allocate its resource.

Our main contribution in this paper is as follows.

- We propose a cost-effective and scalable method of conducting CER via personal health messages.
- We demonstrate this approach is consistent with existing literature in breast cancer and heart disease treatments.
- We propose a precise age and gender extraction approach, showing how these approaches can be used in aggregating users' preference towards treatments of interest.

Our paper is organized as follows. Section 2 details some of the existing work done in CER, data mining and information retrieval. In Section 3, we describe the dataset and keywords that are used in our CER studies. The three studies – author comparison, population comparison, and demographics comparison – are defined in Section 4. An age extraction algorithm is explained in the next section. We then show CER results in Section 6, and conclude and provide future work in Section 7.

2. RELATED WORK

One criterion for CER research is to conduct research on a general population [30] rather than a specialized one that has preexisting symptoms. It is further recommended that an adequate number of all relevant population and patient subgroups be covered so that particular medication can accurately be administered to the group of interest. In our study, we compared various coronary and breast cancer treatments, both covering the effectiveness in population and demographics level. Comparing treatment options for these sicknesses are on the list of important CER research questions to answer [30].

Some comparative studies have been done for both heart disease [9, 11] and breast cancer [42, 14, 16]. Some studies compared the effectiveness of drugs [15, 28, 11, 42, 14] without examining the demographic information of the sample. In our work, these types of studies correspond to population effectiveness study. Some other classes of studies compared drug effectiveness between different demographic groups [11], or compared multiple drugs within a particular group [28, 9]. These allow deciding which medication works best for the given groups of interest. Our approach is also capable of conducting demographic comparison study.

There has been various work using personal health messages to detect health issues. For example, some work utilized Twitter in detecting trends in disease epidemics [31, 5, 33, 13]. This line of work demonstrated how the particular rise and fall of epidemic can be modeled over time or location. Another line of work focused on opinions of various treatments [10, 22], both of which were done in our research group. This work has analyzed how people feel about a particular line of medications over time [10], or summarized what users feel about particular medications [22] based on their aspects. Our current work is different from these as we compare treatments directly both on population and on demographic level. Furthermore, the previous approaches have focused on providing qualitative results via natural language processing techniques while the current work focuses more on quantitative comparative results.

3. PROBLEM SETTING

3.1 Medical Forum Dataset Description

We chose MedHelp[1] posts to conduct our CER study. MedHelp is an online forum where patients and health providers discuss various health-related topics and has over one million health messages. Patients post questions regarding specific health problems or questions, and other patients or experts answer them.

We chose breast cancer and heart disease in conducting CER studies because both of the disease are fairly common and of immediate interest. Over 80 million adults have heart related problems [4] and more than 200,000 American women are diagnosed annually with breast cancer [1].

MedHelp has two major types of forums of interest: support forums and expert forums. Support forums are mostly run by patients who give or seek advise. Expert forums, on the other hand, have certified doctors who give suggestions based on what the patients post. Compared to the support forums, expert forums have less participants per thread. In conducting our research, we used both the support and expert forums. MedHelp allows distinguishing between experts and regular users. We have removed all posts by experts and utilized only those that are written by regular users to comply with the definition of personal health messages. There were 40,996 and 98,644 posts in breast cancer and heart disease forums, respectively. An average user wrote 2.69 and 2.30 posts on each of the two forums. The posts were mostly in the forms of question/advise-seeking and answering.

Because some forums cover similar topics, we are able to aggregate some of these together. We combined 'Breast cancer support forums' and 'Breast cancer expert forums' into the data we ran breast cancer related CER studies on. 'CAD support forums,' 'CHF support forums,' and 'Heart disease support forums' and 'Heart disease expert forums' were combined in running CER studies on heart disease.

3.2 Treatments used in CER study

The treatment classes we investigated on are anticoagulants, blockers, devices, and inhibitors for heart disease, and hormonal, radiation, and chemotherapy treatments for breast cancer. We employed a top-down and a bottom-up approach in collecting keywords that are utilized in each treatment class. For the top-down approach, we used treatment descriptions from Mayo Clinic [2] and WebMD[2] in

[1]http://www.medhelp.org
[2]http://www.webmd.com

the 'treatments' session for breast cancer and heart disease to collect keywords. For the buttom-up approach, we utilized both MetaMap and MedLinePlus Connect to aid us in constructing keywords for each class of treatments. MetaMap [7] is a tool developed at the National Library of Medicine for mapping raw English text to standardized medical concepts in the Unified Medical Language System (UMLS) Metathesaurus. MedlinePlus [3] is the National Institute of Health's Web site for patients. MedlinePlus Connect provides a REST-based service to respond to requests that are queried. Using these tools, we ran MetaMap and collected all the words and phrases that have the semantic type 'phsu' (Pharmacologic substance' in the web forum. We then queried these words to MedLinePlus Connect to see if they contain the medication class types we are interested in. This collection of words is added on top of the keywords that are collected manually. Due to space constraints, we have uploaded the list of keywords on the author's webpage.

4. CER METHODOLOGY

4.1 Definitions

In this section, we will first define CER as it is used throughout this paper. We will then introduce some definitions for the calculation of user preference motivated by [25]. Based on this method, we further extended the sentiment analysis by introducing context sentiment and treatment sentiment.

Effectiveness, in the context of CER, is defined as how well patients respond to given treatments or medications. These are conducted and compared on different demographics under different conditions. We found user preference in web forums to be a useful signal in determining effectiveness. Preference is defined as whether users prefer one particular treatment over another, in terms of the individual author's opinion, population as a whole, or by demographics. Negative preference towards a particular treatment may indicate either the treatment was ineffective or it had numerous side effects. A positive one may indicate the author is content with the outcome. We show empirically for two diseases (heart disease and breast cancer) on numerous treatments in Section 4 that preference is indeed consistent with effectiveness.

We note that users often write about a particular experience in a span of multiple sentences. In order to capture this, we introduce a concept called 'Surroundings' in Definition 1 which captures this intuition.

DEFINITION 1 (SURROUNDINGS). *Given a collection of documents, i-th sentence of d-th document is denoted as $s_{d,i}$. The surroundings for a given sentence $s_{d,i}$ is defined as $E_{d,i} = \{s_{d,i+w},\ \forall w \in [-W, W]\}$, where W indicate the size of context sentences to include. Surrounding at location w from the sentence $s_{d,i}$ is defined as $\epsilon_{d,i,w} \in E_{d,i}$.*

As an illustrative example, surrounding of sentence $s_{d,i}$ with $W = 1$ includes sentences $s_{d,i-1}, s_{d,i}$, and $s_{d,i+1}$. We noted that the naive definition of surroundings is insufficient to explain the user's experience. Oftentimes, users may talk about a particular concept in a given sentence, and switch to talk about a different concept in the next. Context, more rigorously defined in Definition 2, address this concern by removing any sentences that included concepts different from our target concept.

DEFINITION 2 (CONTEXT). *Let treatment $t_{s_{d,i}} \in T$ be contained in sentence $s_{d,i}$. Given treatment $t_{\epsilon_{d,i,w}}$ that is contained in $\epsilon_{d,i,w} \in E_{d,i}$, the surrounding of $s_{d,i}$ is consistent if $t_{\epsilon_{d,i,w}} - t_{\epsilon_{d,i,0}} = \phi,\ \forall w \in [-W, W]$. Context $c_{d,i}$ is defined as the union of all the surroundings with treatment $t_{\epsilon_{d,i,w}}$ that is consistent.*

For each of the contexts $c_{d,i}$, we ran sentiment analysis tool[3] to obtain sentiment scores. The tool outputs positive sentiment score, $a_{p,c_{d,i}}$ and negative sentiment score, $a_{n,c_{d,i}}$ of the given context. The attitude of the context is defined by combining the positive and negative sentiment scores, more rigorously defined in Definition 3.

DEFINITION 3 (ATTITUDE OF THE CONTEXT). *Given positive and negative sentiment outputs, $a_{p,c_{d,i}} > 0$ and $a_{n,c_{d,i}} > 0$, of the context $c_{d,i}$, the attitude of the given context is defined as $a_{c_{d,i}} = a_{p,c_{d,i}} - a_{n,c_{d,i}}$. Each of $a_{c_{d,i}} > 0$, $a_{c_{d,i}} < 0$ and $a_{c_{d,i}} = 0$ refer to positive, negative, and neutral sentiments, respectively.*

In order to extract a user's opinion towards treatments t, we collect all the attitudes toward the treatment for the user and take the arithmetic mean, more formally defined in Definition 4.

DEFINITION 4 (USER'S ATTITUDE TOWARDS TREATMENT). *Given the user u we wish to extract the opinion from, let us denote the set of all the contexts of treatment $t \in T$ by the user u be denoted as A_t^u. The user's final attitude towards the treatment t is defined as $a_t^u = \sum_{a_{c_{d,i}} \in A_t^u} a_{c_{d,i}}$.*

Note that in this analysis, we examined only the valence of attitudes but not the magnitude by using sentiment (positive, negative, and neutral) to classify the user's opinion as defined in Definition 3. Future research may further explore this issue.

4.2 Author Effectiveness Comparison

Author effectiveness comparison compares treatments that are mentioned by the same author. Oftentimes patients have experience with multiple medications. If the author had positive experience with a particular treatment after exhausting all the other options, their preference will reflect this. On the other hand, if all the treatments they have taken had similar efficacies, they will not have a preference on one treatment over another. User preference is used to determine whether a particular user prefers treatment A over treatment B, defined in Definition 5.

DEFINITION 5 (USER PREFERENCE). *Given a user, u, treatments t, t', $t \neq t'$, and the person's attitude towards treatments, a_t^u and $a_{t'}^u$ as is computed by Definition 4, we say treatment t is preferred over t' for user u if $sgn(a_t^u) > sgn(a_{t'}^u)$. There is no preference between treatment t to t' if $sgn(a_t^u) = sgn(a_{t'}^u)$.*

It is important to note from the definition that each of the users are weighted equally to prevent active users from overwhelming treatment preferences.

Given user preference, we can now define treatment preference based on users that compare two different treatments. Treatment preference shows, for patients who have been exposed to two different treatments, which one they prefer and is defined in Definition 6.

[3]We used LIWC to analyze sentiment.

DEFINITION 6 (TREATMENT PREFERENCE). *Let us denote $A^U_{t,t'}$ as set of attitudes the set of users U have towards treatments t and t'. Then, the number of users who prefer treatment t is given by $|U_t|$. If $|U_t|$ is statistically significantly larger than $|U_{t'}|$ then treatment t is preferred over t', and vice versa. Otherwise, there is no preference.*

4.3 Population Effectiveness Comparison Study

In some medical studies, two or multiple demographic groups are sampled and analyzed to compare their preferences on the target medicines.

Similarly, we compare the treatments by the percentages of people in specific demographic groups (e.g., male, female, young, old) with positive or negative opinions. We define this approach to compare treatments as a population effectiveness study. Specifically, we conducted two forms of comparison: within group comparison by comparing the preferences on multiple treatments for patients belonged to one specific demographic group, and cross-group comparison by comparing opinions of patients from different demographic group on one specific treatment.

To conduct demographic comparison, we used results from the population effectiveness comparison instead of author effectiveness comparison because the latter generates too sparse data in this particular dataset. In fact, since the distribution of the online user's posting behavior follows a long tail distribution [25], most users participate in limited number of threads and therefore have low likelihood of mentioning multiple treatments. To enable valid statistical analysis with sufficient power, we decided to use only population results. Future research may explore the other method when a large dataset can be obtained, possibly be aggregating multiple resources.

Population effectiveness studies are defined as the following. Each user u has sentiment towards a treatments $t \in T$ as mentioned in Definition 4. We count the number of users with each category of sentiment (positive, negative, neutral) towards the treatment. For each treatment, we then compute the proportion of people with each sentiment over the total number of people who mentioned the particular treatment. Finally, we conduct statistical analysis to compare the proportion of positive and negative opinions on pairs of treatments. Specifically, if a higher proportion of positive opinions are given to one over the other, the former is generally preferred over the latter. On the other hand, if a lower proportion of negative opinions are given to one over the other, the former might be less effective or have more side effects than the latter. We consider treatment A is preferred over treatment B only when A has a significantly higher proportion of positive opinions and significantly lower proportion of negative opinions. Finally, because this is a population study on the effectiveness of treatments, we assigned the same weights for all users who poss personal health messages.

4.4 Demographics Effectiveness Comparison

We compare demographic preference by utilizing population effectiveness methodology described in the previous section. In particular, for each of the subset of demographics we are interested in, we conduct methodology identical to those conducted in population effectiveness study for both within group and between group comparison (for the latter, the proportion is compared for different groups instead of different treatments). The demographic comparison we are interested in are gender (male, female) and age (20-44, and 45 and up). We are able to extract age and gender information by leveraging publicly available demographics information on MedHelp user profile pages. We further extract age information not on user profile pages by introducing a supervised, rule-based method which is explained in detail in the next section.

In comparing treatments for demographics class, we wish to answer the following questions:
1) What treatments are effective for given demographics?
2) Which demographics prefer particular treatment compared to the other?
We define the study that answers 1) as within-group analysis, and that of 2) as cross-group analysis.

5. DEMOGRAPHIC EXTRACTION

Many web forums do not have publicly available demographic information in user profiles, and for those that do, many users choose not to publish such information. Therefore, CER research on demographic effectiveness comparison may need to rely on text processing to extract this information from user post content. This further allows us to aggregate results from users we were previously unable to extract age information from, increasing granularity of demographic effectiveness comparison results.

Previous literature on demographic extraction can be clustered into two themes. The first involves traditional supervised learning such as using linear regression or SVM. These approaches take features from given text [6, 29] or search log patterns [20]. These methods allow predicting the user's age for all the users. The second approach uses rule-based approaches [43] in extracting user's demographics. These methods require researchers to hard-code rules that indicate demographic information. These top-down approaches tend to have high precision but are limited by the researchers' ability to write down rules, limiting its recall. In general, the first approach is preferred when an approximate estimate of demographics information is acceptable while the second approach is used when high precision is required.

We propose a hybrid method that utilizes supervision in generating rules for user demographics extraction. While deploying a traditional supervised approach would have allowed better coverage, we used rule generation approach because we are mainly concerned with ensuring high precision, and we did not want mentions of treatments or medications to potentially influence age classification results.

The hybrid approach first collects potential phrases that contain age information. It takes as input tuple (D, U), which corresponds to collection of personal health messages, D and the author of the corresponding messages, U, collection of user profiles U_p (which contain publicly available demographics information), and α and β, parameters which guide how strict the inferred rules should be. U_p serves as labels to be used for demographics information of interest. Any phrase that contains a number that matches the age in the user's profile page is defined as a potential phrase, and we label them as *phrase* in Algorithm 1, set of *phrase* as *phrases*. These potential phrases are limited to $n+1$ words, with $\frac{n}{2}$ words in front and back of the number that has been matched. It is important to note that all the numerics in potential phrases are replaced with a numerics token. These steps refer to the first for-loop on Algorithm 1.

Algorithm 1 Age Phrase Learner

Input: (D, U), U_p, α, β
Output: $fsp, mfsp$
  $phrases = \{\}$
  for $(d, u) \in (D, U)$ **do**
    if $u_p \in U_p$ **then**
      $phrase \leftarrow getPhrases(d, age(u_p))$
      $phrases = phrases \cup phrase$
    end if
  end for
  $fsp \leftarrow PrefixSpan(phrases)$
  $mfsp = \{\}$
  while fsp not changed **do**
    for $u \in U$, $d \in docs(u)$, $s \in d$ **do**
      $(Age, Pattern) \leftarrow getAge(s, fsp, mfsp, dwords)$
      $updatePatternPrecision(Pattern)$
    end for
    for $Pattern \in fsp$ below β precision **do**
      Remove $Pattern$ from fsp
      if $Pattern$ below α precision **then**
        add $Pattern$ to $mfsp$
      end if
    end for
  end while

Frequent sequence pattern mining algorithm called PrefixSpan [32] is then run on *phrases*. The algorithm is both complete and efficient in that it can find all possible frequent sequences, and is fast enough for our purpose. We label the frequent sequences that are returned by PrefixSpan as 'fsp.' Using merely the frequent sequence patterns, however, often results in a high misclassification rate. Some of the frequent patterns may actually refer to some other quantitative property (such as dosage of medication) but it so happens that this number matches the person's age. To alleviate the problem, we also collect frequently misclassified phrases. These are phrases which have precision below a certain threshold α. We label frequently misclassified phrases as 'mfsp.' Frequent patterns are removed if they are below β. These two parameters are used to control two facets of the algorithm. Higher β denotes how aggressive we wish to be in removing patterns whereas higher α indicates how actively we wish to avoid seeing particular patterns. To expedite the convergence speed, we have manually labeled a few words and suffixes that do not indicate age. These words are *week, weeks, day, days, month, months, lb, lbs, pound, pounds, %, mg, kg, ml*, and when these patterns are seen in a phrase, we disregarded the match. The collection of these words correspond to 'dwords' in the algorithm. Age is extracted from the sentence if it matches a phrase in fsp but not in $mfsp$. In Algorithm 1, this corresponds to the *getAge* function. We repeat the process until fsp converges.

Age inference on the user is done using the learned phrases. For each user, all the sentences are tested for a match in fsp but not in $mfsp$. If this test succeeds, then the age is extracted using the frequent pattern that has been used. In the case of ambiguity, we take the mode of the matched phrases. We note the same procedure can be used in gender extraction by matching gender related words as opposed to its age. This is done by giving initial set of keywords that represent each gender and then learning frequent patterns. Due to its similarity with the age extraction algorithm, we do not include the gender extraction algorithm in this paper. Age inference is shown in Algorithm 2.

Runs	A	B	C	D	E	F
Baseline	15,234	3,726	9,111	6,892	2,219	0.47
Our method	15,234	3,726	4,855	3,668	1,159	0.94
Baseline	42,860	10,634	31,189	23,602	7,587	0.40
Our method	42,860	10,634	19,157	14,921	4,236	0.84

Table 1: First two rows : results from breast cancer forums, last two rows : results from heart disease forums, A : #Users total in forum, B : #Users with known age information, C : #Users' age information inferred, D : #Users' age information inferred but not listed on the profile page, E : #Users inferred & has age info, F : Precision ($\frac{Correct\ \&\ E}{E}$)

Disease	Breast Cancer Treatments		
Treatment 1 (T1) Treatment 2 (T2)	Radiation Hormonal	Chemotherapy Hormonal	Chemotherapy Radiation
N1 (Prefer T1 to T2)	205	209	232
N2 (Prefer T2 to T1)	160	161	206
χ^2	36.17	23.20	1.54
Degree of Freedom	1	1	1
p-value	< 0.01	< 0.01	0.21
$T1 > T2$?	Yes	Yes	No

Table 2: Author comparison on Breast Cancer

Algorithm 2 Age Inference

Input: $(D, U), fsp, mfsp, dwords$
Output: $\{(u, age)\}$
  $\{(U, Ages)\} \leftarrow \{\}$
  for $u \in U$ **do**
    $potAges = \{\}$
    for $d \in docs(u)$, $s \in d$ **do**
      $(age, Pattern) \leftarrow getAge(s, fsp, mfsp, dwords)$
      $potAges \leftarrow potAges \cup age$
    end for
    $\{(U, Ages)\} \leftarrow \{(U, Ages)\} \cup \{(u, mode(potAges))\}$
  end for

The parameters we set for the algorithm are $\alpha = 0.15$, $\beta = 0.5$, $n = 10$. We compared our method with a baseline rule-extraction approach and the results are summarized on Table 1. The baseline approach does not remove patterns with high misclassification rate, and not surprisingly achieves low precision (0.47 in breast cancer forums, 0.4 in heart forums). Our algorithm has precision of 0.94 and 0.84 on breast cancer forums and heart forums, respectively. We were also able to infer approximately 30% of users' age information. Combining both the users with inferred demographic information and those whose information is listed on the profile page, we were able to double the amount of users' demographic information available.

6. CER STUDIES

In this part, we present results from three experiments: author effectiveness comparison, population effectiveness comparison and demographics effectiveness comparison on two diseases: heart disease, and breast cancer. For heart disease, four categories of treatments were compared: device, anticoagulants, inhibitors and blockers. For breast cancer, three categories of treatments were compared: chemotherapy, ra-

diation, and hormonal treatments. All of them are common treatments for the corresponding disease and are discussed frequently on the forum. There were a total of 15,234 users and 65,853 health messages on breast cancer forums, and 42,860 users and 187,155 health messages on heart disease forums.

6.1 Author Effectiveness Comparison

In the first experiment, we compared each pair of treatments based on the opinions of all authors who have ever commented on both treatments. Specifically, when comparing treatment A and treatment B, we counted how many authors prefer A to B (corresponding to N1), and how many prefer B to A (corresponding to N2), then we conducted Chi-square tests on the two numbers to identify pairs of comparisons for which significantly more people preferred one way over the other.

Table 2 and 3 show the results of each pair of comparison treatments on heart disease and breast cancer respectively. For heart disease, it suggests that people expressed more positive opinions on anticoagulants and devices than blockers and inhibitors for treating heart diseases. Also people were significantly more positive on blockers than inhibitors. There is no significant difference in the preference between anticoagulants and devices observed. For breast cancer treatments, people expressed more positive opinions on radiation and chemotherapy than hormonal treatments. We did not observe significant difference in the preference between chemotherapy and radiation.

6.2 Population Effectiveness Comparison

In this experiment, we compared each pair of treatments based on the overall opinion sentiment, i.e., we included all the comments on each treatment and compared the proportions of people who showed positive and negative sentiment between each pair of treatments. We used an independent two-sample proportional test to identify pairs for which a significantly higher proportion of people expressed positive or negative opinions on one treatment respectively. To make the test more rigorous, we only consider pairs for which significantly ($p < 0.05$) or marginally significantly ($0.05 \leq p < 0.10$) higher proportion of people expressed positive opinions, meanwhile significantly or marginally significantly lower proportions of people expressed negative opinions on one treatment to qualify a valid difference. That is, if significantly more people expressed positive opinions on treatment A than treatment B, while significantly more people also expressed negative opinions on A than B, we consider it to be controversial, and thus cannot conclude which treatment is preferred.

Table 4 and 5 show the results of indirect comparison for heart disease and breast cancer. Consistent with the results of author comparison, it indicates that higher proportion of patients expressed positive opinions on anticoagulants and devices than inhibitors and blockers in treating heart disease. While consistent with author comparison, we found significantly a higher proportion of people showed positive opinions on blockers than inhibitors, however, we did not observe there are significantly lower proportion of people that showed negative opinions on blockers than inhibitors. We therefore cannot conclude the preference between blockers and inhibitors. For treating breast cancer, people are more likely to prefer radiation and chemotherapy to hormonal treatments.

Disease	Breast Cancer Treatments		
Treatment 1 (T1) Treatment 2 (T2)	Radiation Hormonal	Chemotherapy Hormonal	Chemotherapy Radiation
N1	2,393	2,878	3,878
N2	1,680	1,680	2,393
p(positive on T1)	0.41	0.44	0.44
p(positive on T2)	0.39	0.39	0.41
χ^2 (Positive)	3.46	12.20	3.00
Degree of Freedom	1	1	1
p-value (Positive)	0.05	< 0.01	0.09
p(negative on T1)	0.26	0.29	0.29
p(negative on T2)	0.34	0.34	0.26
χ^2 (Negative)	29.42	14.78	3.93
Degree of Freedom	1	1	1
p-value (Negative)	< 0.01	< 0.01	0.05 (reverse)
$T1 > T2$?	Yes	Yes	No

Table 5: Population comparison on Breast Cancer

Disease	Breast Cancer Treatments		
Treatment 1 (T1) Treatment 2 (T2)	Radiation Hormonal	Chemotherapy Hormonal	Chemotherapy Radiation
N1	739	770	770
N2	525	525	739
p(positive on T1)	0.46	0.46	0.46
p(positive on T2)	0.41	0.41	0.46
χ^2 (Positive)	2.98	3.66	0.02
Degree of Freedom	1	1	1
p-value (Positive)	0.08	0.05	0.89
p(negative on T1)	0.24	0.29	0.29
p(negative on T2)	0.34	0.34	0.24
χ^2 (Negative)	16.57	4.09	4.90
Degree of Freedom	1	1	1
p-value (Negative)	< 0.01	0.04	0.03 (reverse)
$T1 > T2$?	Yes	Yes	No

Table 6: Demographic comparison for Breast Cancer on older population

6.3 Demographics Effectiveness Comparison

6.3.1 Age Analysis

Our approach of mining online health messages can potentially generate preferential results between different demographic groups, which could be valuable for supporting CER. With the current demographic information extraction method, we are able to obtain 7,394 users' age and 6,465 users' gender information from breast cancer forums and 25,555 users' age and 17,618 users' gender information from heart disease forums. It is important to keep in mind we cannot use all of these users in our study as not all users mention treatments that we are interested in conducting CER study. The demographics information enabled us to perform:
1) within group analysis, by identifying any pair of comparison that shows inconsistent results with the overall trend.
2) cross group analysis, by identifying significant preferential difference between two demographic groups on any single treatment.
Specifically, we analyzed between older group (45 and above), and younger (age between 20 and 44), and between female and male groups. Similar to population effectiveness comparison experiment, we used an independent two-sample proportional test to conduct the analysis.

For breast cancer, we only analyzed the age group since it is dominated by female patients. In this section we will first report the results of analysis performed on younger and older age group for both heart disease and breast cancer,

Disease	Heart Disease Treatments					
Treatment 1 (T1) Treatment 2 (T2)	Anticoagulants Inhibitor	Anticoagulants Blocker	Anticoagulants Device	Device Inhibitor	Device Blocker	Blocker Inhibitor
N1 (Prefer T1 to T2)	154	256	140	148	261	261
N2 (Prefer T2 to T1)	65	158	126	70	187	204
χ^2	36.17	23.20	0.74	27.91	12.22	6.99
Degree of Freedom	1	1	1	1	1	1
p-value	< 0.01	< 0.01	0.39	< 0.01	< 0.01	0.01
$T1 > T2$?	Yes	Yes	No	Yes	Yes	Yes

Table 3: Author comparison on Heart Disease

Disease	Heart Disease Treatments					
Treatment 1 (T1) Treatment 2 (T2)	Anticoagulants Inhibitor	Anticoagulants Blocker	Anticoagulants Device	Device Inhibitor	Device Blocker	Blocker Inhibitor
N1	2,162	2,162	2,162	2,457	2,457	7,257
N2	2,422	7,257	2,457	2,422	7,257	2,422
p(positive on T1)	0.35	0.35	0.35	0.34	0.34	0.27
p(positive on T2)	0.22	0.27	0.34	0.22	0.27	0.22
χ^2 (Positive)	36.17	23.20	0.74	27.91	12.22	6.99
Degree of Freedom	1	1	1	1	1	1
p-value (Positive)	< 0.01	< 0.01	0.39	< 0.01	< 0.01	0.01
p(negative on T1)	0.40	0.40	0.40	0.43	0.43	0.54
p(negative on T2)	0.56	0.54	0.43	0.56	0.54	0.56
χ^2 (Negative)	110.60	130.55	5.34	75.95	82.18	2.61
Degree of Freedom	1	1	1	1	1	1
p-value (Negative)	< 0.01	< 0.01	0.02	< 0.01	< 0.01	0.11
$T1 > T2$?	Yes	Yes	No	Yes	Yes	No

Table 4: Population comparison on Heart Disease

and then report the analysis on the female and male group for heart disease.

Table 6 and 8 show the within-group comparison results of the older group. For both heart disease and breast cancer, older adults exhibited consistent preference with the overall trend (see Table 4 and 5). That is, a significantly higher proportion of older patients had positive opinions on anticoagulants and devices than on inhibitors and blockers. For breast cancer, a higher proportion of older patients had more positive opinions on chemotherapy and radiation than hormonal treatment.

Table 7 and 9 show the within-group comparison results for younger group. We found three pairs of comparison that were significant in the overall analysis become non-significant in the younger group: device v.s. inhibitor, radiation v.s. hormonal and chemotherapy v.s. hormonal treatment. Closer observation revealed that for hormonal treatments, no difference in proportion of younger adults and older adults expressed negative opinions was observed. This set of results indicates that younger people do not necessarily favor devices over inhibitors for treating heart disease, and they do not have more negative opinions on hormonal treatments than radiation or chemotherapy.

6.3.2 Gender Analysis

We now look at the within and cross-group analysis on male and female patients for heart related disease. Table 10 shows the results for female patients. It indicates that, consistent with the overall tendency, a higher proportion of female patients prefer anticoagulants and devices to blockers and inhibitors and is shown in Table 10.

Table 11 shows the result for male patients. We found two pairs of comparison show different results from the overall trend and female group: 1) the comparison between anticoagulants and blockers does not show significantly preferential result; 2) male patients are more likely to prefer blockers over inhibitors. The data showed preference results that are worth further study in clinical trials.

Disease	Breast Cancer Treatments		
Treatment 1 (T1) Treatment 2 (T2)	Radiation Hormonal	Chemotherapy Hormonal	Chemotherapy Radiation
N1	310	407	407
N2	273	273	310
p(positive on T1)	0.50	0.49	0.49
p(positive on T2)	0.42	0.42	0.50
χ^2 (Positive)	3.04	2.78	0.02
Degree of Freedom	1	1	1
p-value (Positive)	0.08	0.10	0.89
p(negative on T1)	0.26	0.27	0.27
p(negative on T2)	0.28	0.28	0.26
χ^2 (Negative)	16.57	4.09	4.90
Degree of Freedom	1	1	1
p-value (Negative)	0.71	0.88	0.85
$T1 > T2$?	No	No	No

Table 7: Demographic comparison on Breast Cancer on younger population

6.3.3 Cross-group Analysis

For cross-group analysis, we identify treatments that have significantly different proportions of older and younger adults that expressed either positive or negative opinions. Specifically, we found the following interesting results: For blockers, a lower proportion of older people expressed negative opinions than younger ones did ($\chi^2 = 3.42,\ p = 0.06$). For hormonal treatment, a lower proportion of younger people expressed negative opinions than older ones did ($\chi^2 = 3.14,\ p = 0.08$). We did not observe any statistically significant results in cross-group analysis for female and male patients' opinions on treatment for heart disease.

6.4 Case Studies

To validate the results of our experiments, we searched exisiting literature for relevant evidence. We found at least

Disease	Heart Disease Treatments					
Treatment 1 (T1) Treatment 2 (T2)	Anticoagulants Inhibitor	Anticoagulants Blocker	Anticoagulants Device	Device Inhibitor	Device Blocker	Blocker Inhibitor
N1	153	153	153	166	166	414
N2	217	414	166	217	414	217
p(positive on T1)	0.35	0.35	0.35	0.39	0.39	0.26
p(positive on T2)	0.23	0.26	0.39	0.23	0.26	0.23
χ^2 (Positive)	5.95	4.08	0.58	11.59	9.83	0.54
Degree of Freedom	1	1	1	1	1	1
p-value (Positive)	0.01	0.04	0.45	< 0.01	< 0.01	0.46
p(negative on T1)	0.41	0.41	0.41	0.45	0.45	0.56
p(negative on T2)	0.62	0.56	0.45	0.62	0.56	0.62
χ^2 (Negative)	15.39	9.51	0.38	10.49	5.29	1.99
Degree of Freedom	1	1	1	1	1	1
p-value (Negative)	< 0.01	< 0.01	0.53	< 0.01	0.02	0.16
$T1 > T2$?	Yes	Yes	No	Yes	Yes	No

Table 8: Demographic comparison on Heart Disease on older population

Disease	Heart Disease Treatments					
Treatment 1 (T1) Treatment 2 (T2)	Anticoagulants Inhibitor	Anticoagulants Blocker	Anticoagulants Device	Device Inhibitor	Device Blocker	Blocker Inhibitor
N1	75	75	75	106	106	374
N2	120	374	106	120	374	120
p(positive on T1)	0.39	0.39	0.39	0.31	0.31	0.21
p(positive on T2)	0.23	0.21	0.31	0.23	0.21	0.23
χ^2 (Positive)	5.13	9.96	0.80	1.73	4.35	0.07
Degree of Freedom	1	1	1	1	1	1
p-value (Positive)	0.02	< 0.01	0.37	0.19	0.04	0.80
p(negative on T1)	0.43	0.43	0.43	0.45	0.45	0.62
p(negative on T2)	0.58	0.62	0.45	0.58	0.62	0.58
χ^2 (Negative)	15.39	9.51	0.38	10.49	5.29	1.99
Degree of Freedom	1	1	1	1	1	1
p-value (Negative)	0.06	< 0.01	0.84	0.09	< 0.01	0.41
$T1 > T2$?	Yes	Yes	No	No	Yes	No

Table 9: Demographic comparison on Heart Disease on younger population

Disease	Heart Disease Treatments					
Treatment 1 (T1) Treatment 2 (T2)	Anticoagulants Inhibitor	Anticoagulants Blocker	Anticoagulants Device	Device Inhibitor	Device Blocker	Blocker Inhibitor
N1	366	366	366	511	511	1,337
N2	418	1,337	511	418	1,337	418
p(positive on T1)	0.35	0.35	0.35	0.33	0.33	0.23
p(positive on T2)	0.21	0.23	0.33	0.21	0.23	0.21
χ^2 (Positive)	17.58	19.64	0.24	15.52	9.83	0.66
Degree of Freedom	1	1	1	1	1	1
p-value (Positive)	< 0.01	< 0.01	0.62	< 0.01	< 0.01	0.42
p(negative on T1)	0.41	0.41	0.41	0.44	0.44	0.56
p(negative on T2)	0.56	0.56	0.44	0.56	0.56	0.56
χ^2 (Negative)	18.74	27.75	0.63	14.61	23.62	0
Degree of Freedom	1	1	1	1	1	1
p-value (Negative)	< 0.01	< 0.01	0.43	< 0.01	< 0.01	1
$T1 > T2$?	Yes	Yes	No	Yes	Yes	No

Table 10: Demographic comparison on Heart Disease on female

Disease	Heart Disease Treatments					
Treatment 1 (T1) Treatment 2 (T2)	Anticoagulants Inhibitor	Anticoagulants Blocker	Anticoagulants Device	Device Inhibitor	Device Blocker	Blocker Inhibitor
N1	351	351	351	397	397	1,001
N2	469	1,001	397	469	1,001	469
p(positive on T1)	0.30	0.30	0.30	0.35	0.35	0.25
p(positive on T2)	0.20	0.25	0.35	0.20	0.25	0.20
χ^2 (Positive)	8.65	2.31	1.81	20.84	11.60	3.82
Degree of Freedom	1	1	1	1	1	1
p-value (Positive)	< 0.01	0.13	0.18	< 0.01	< 0.01	0.05
p(negative on T1)	0.46	0.46	0.46	0.46	0.46	0.53
p(negative on T2)	0.59	0.53	0.46	0.59	0.53	0.59
χ^2 (Negative)	13.21	5.96	0.04	13.07	5.70	3.40
Degree of Freedom	1	1	1	1	1	1
p-value (Negative)	< 0.01	0.01	0.95	< 0.01	< 0.01	0.11
$T1 > T2$?	Yes	No	No	Yes	Yes	Yes

Table 11: Demographic comparison on Heart Disease on male

eight of our findings are supported by medical works. The rest, for which we found few relevant literatures (neither proof nor disproof), could be explored by future medical research.

We divide the study into two parts for both heart disease and breast cancer treatments: population effectiveness comparison and demographics effectiveness comparison. The former compares why the people prefer treatment A over the other, or indirect explanation of these comparison. The latter compares demographic differences or similarities for given treatment.

6.4.1 Population Effectiveness Comparison

Case 1. *Chemotherapy v.s. Hormonal therapy*:
Our study showed that patients had more positive opinions on chemotherapy than hormonal therapy in treating breast cancer. A paper published in Cochrane Reviews, which involved 7 clinical trials and more than 700 patients concluded that chemotherapy is advantageous over hormonal therapy in reducing the tumor response rate [41]. This is consistent with our results.

Case 2. *Radiation v.s. Hormonal therapy*:
Our study showed that patients favor radiation therapy over hormonal therapy in treating breast cancer. It seems to be consistent with what previous medical research concluded. Specifically, it was found that breast cancer patients who had radiation therapy showed lower post-treatment side effects than hormonal and combinational treatments [42, 14].

Case 3. *β Blockers v.s. ACE Inhibitors*:
Our study showed blockers are preferred over inhibitors in author comparison, and blockers had significantly more positive sentiment over inhibitors in population studies. Two works on heart failure claim that β blockers are as effective as ACE inhibitors alone [35]. In some other cases, ACE inhibitors [36] work better than β blockers which is consistent with our findings.

Case 4. *Device v.s. one of β blockers or inhibitors*:
Devices were significantly preferred over inhibitors or blockers from our studies. One work concludes that a combination of device (Left Ventricular Assist Device) and pharmaceutical treatments such as ACE inhibitors or β blockers were more effective than using pharmaceutical treatments alone in treating heart failure [8]. Another study showed treatment effects in LVAD were four times more than that of β blockers and ACE inhibitors in preventing death of end stage heart failure patients [18]. Finally, one study showed using prophylactic pacemakers facilitated β blocker treatment [39], further providing reasons why these devices may be effective.

Case 5. *Anticoagulants, devices and β blockers*:
Our study consistently showed there is no preference between anticoagulants and devices. Many patients requiring pacemakers or implantable cardioverter-defibrillator (ICD) surgery take warfarin [9], which is a very commonly used anticoagulants. This can be why there was no preference for anticoagulants over devices and explains why anticoagulants were preferred over blockers and inhibitors. Furthermore, warfarin appear to be at least as, if not more, effective as β blockers in reducing reinfarction rates compared to placebo pills [37, 24] but warfarin is more cost effective than β blockers, another contributing factor why anticoagulants may be preferred over β blockers.

6.4.2 Demographic Effectiveness Comparison

Case 1. *Effects of ACE inhibitor by gender*:
Our results indicate there is no differences in ACE inhibitor based on gender. A study sponsored by Agency for Healthcare Research and Quality indicates ACE inhibitors reduce composite efficacy endpoints similarly in males and females, which is consistent with our findings [11].

Case 2. *Comparison of β blockers by age*:
Our results indicate older people were less negative towards β blockers than younger people. One observational study, which had cohort size of over 10,000, concludes that efficacy of β blockers seem to be extend to elderly [17], and β blocker seem to be dependent on the dosage. However, younger people have trends of being more impacted by cognitive impairment than older people [21], which may explain why younger people were more negative about β blockers than older people.

Case 3. *Hormonal v.s. Chemotherapy on older group*:
Our results indicate that chemotherapy is significantly preferred over hormonal therapy on older population. A review consisting of 133 trials and 75,000 women [15] concluded that between ages 50 and 69, chemotherapy plus tamoxifen (a common hormonal therapy) is better than chemotherapy alone both for recurrence and for mortality. However, chemotherapy was still better than tamoxifen alone in terms of breast cancer recurrence, which is consistent with our results.

7. CONCLUSION AND FUTURE WORK

In this paper, we utilized online personal health messages in conducting CER. Specifically, we aggregated and analyzed messages on medical forums to compare patients' opinions on different treatments for heart disease and breast cancer. We analyzed the preferences for both the overall population as well as different age and gender groups. The demographics of the populations were extracted by utilizing both publicly available user profiles and our high precision demographic information extraction algorithm.

It should be noted that web forums may not be representative of the population as a whole – for example, those that are suffering from terminal illness are unlikely to post on web forums. Despite these weaknesses however, we have shown personal health messages are useful in hypothesis generation. Indeed, we are able to validate many of the comparative effectiveness results with existing literature.

There are various ways this work can further be extended. First, we have only explored one aspect of effectiveness by examining users' preference. It would be beneficial to further investigate user preference based on specific medical aspects, such as side effects or efficacies of the treatments. Methods such as those used in summarizing opinions of given drugs [22] can be extended to such studies. Second, we have provided potential in using forums that do not have publicly available demographical information by introducing our high precision demographic extraction algorithm. This can be utilized to conduct CER by aggregating multiple sources of personal health messages. Finally, we have conducted CER based on their demographics. As a natural extension, a system that conducts CER based on particular symptoms of interest may further aid researchers and practitioners in deciding which treatments are suitable given patient conditions. Exploiting entity relation semantics [19] is a possible

direction that can be further utilized to find symptoms and treatments of interest.

8. ACKNOWLEDGEMENTS

We thank the annonymous reviewers for their insightful comments. This research was supported in part by Health Information Technology Center (HITC) Fellowship at the University of Illinois, Urbana-Champaign and State Farm Doctoral Scholarship. We would also like to thank Sean Massung for helping the authors with the revision.

9. REFERENCES

[1] American Cancer Society. http://www.cancer.org.
[2] Mayo Clinic. http://www.mayoclinic.com.
[3] MedLine Plus. http://www.nlm.nih.gov/medlineplus/.
[4] Million Hearts. http://millionhearts.hhs.gov.
[5] E. Aramaki, S. Maskawa, and M. Morita. Twitter catches the flu: detecting influenza epidemics using twitter. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '11, pages 1568–1576, Stroudsburg, PA, USA, 2011. Association for Computational Linguistics.
[6] S. Argamon and et al. Automatically profiling the author of an anonymous text. *Commun. ACM*, 52(2):119–123, Feb. 2009.
[7] A. R. Aronson and F. M. Lang. An overview of MetaMap: historical perspective and recent advances. *J Am Med Inform Assoc*, 17(3):229–236, 2010.
[8] E. J. Birks and et al. Left ventricular assist device and drug therapy for the reversal of heart failure. *N. Engl. J. Med.*, 355(18):1873–1884, Nov 2006.
[9] D. H. Birnie and et al. Pacemaker or defibrillator surgery without interruption of anticoagulation. *N. Engl. J. Med.*, 368(22):2084–2093, May 2013.
[10] B. W. Chee, R. Berlin, and B. Schatz. Predicting adverse drug events from personal health messages. *AMIA Annu Symp Proc*, 2011:217–226, 2011.
[11] C. CI and et al. Effectiveness of angiotensin converting enzyme inhibitors or angiotensin ii receptor blockers added to standard medical therapy for treating stable ischemic heart disease. 2009.
[12] J. Concato and et al. Observational methods in comparative effectiveness research. *Am. J. Med.*, 123(12 Suppl 1):16–23, Dec 2010.
[13] A. Culotta. Towards detecting influenza epidemics by analyzing twitter messages. In *Proceedings of the First Workshop on Social Media Analytics*, SOMA '10, pages 115–122, New York, NY, USA, 2010. ACM.
[14] R. E. Curtis and et al. Risk of leukemia after chemotherapy and radiation treatment for breast cancer. *N. Engl. J. Med.*, 326(26):1745–1751, 1992.
[15] A. De Schryver, J. Huys, and L. Vakaet. Systemic treatment of early breast-cancer by hormonal, cytotoxic, or immune therapy: 133 randomized trials involving 31000 recurrences and 24000 deaths among 75000 women: 1. *LANCET*, 339(8784):1–15, 1992.
[16] A. Drapkin and C. Merskey. Anticoagulant therapy after acute myocardial infarction. Relation of therapeutic benefit to patient's age, sex, and severity of infarction. *JAMA*, 222(5):541–548, Oct 1972.
[17] B. R. Dulin, S. J. Haas, W. T. Abraham, and H. Krum. Do elderly systolic heart failure patients benefit from beta blockers to the same extent as the non-elderly? Meta-analysis of >12,000 patients in large-scale clinical trials. *Am. J. Cardiol.*, 95(7):896–898, Apr 2005.
[18] M. Eric A. Rose and et al. Long-Term Use of a Left Ventricular Assist Device for End-Stage Heart Failure. *N Engl J Med*, 345(20):1435–1443, Nov. 2001.
[19] X. He, Y. Li, R. Khetani, B. Sanders, Y. Lu, X. Ling, C. Zhai, and B. Schatz. BSQA: integrated text mining using entity relation semantics extracted from biological literature of insects. *Nucleic Acids Res.*, 38(Web Server issue):W175–181, Jul 2010.
[20] J. Hu and et al. Demographic prediction based on user's browsing behavior. In *Proceedings of the 16th international conference on World Wide Web*, WWW '07, pages 151–160, New York, NY, USA, 2007. ACM.
[21] D. JE and N. RP. Neuropsychological side effects of β-blockers. *Archives of Internal Medicine*, 149(3):514–525, 1989.
[22] Y. Jiang, Q. V. Liao, Q. Cheng, R. B. Berlin, and B. R. Schatz. Designing and evaluating a clustering system for organizing and integrating patient drug outcomes in personal health messages. *AMIA Annu Symp Proc*, 2012:417–426, 2012.
[23] K. K. Kim and et al. Development of a Privacy and Security Policy Framework for a Multistate Comparative Effectiveness Research Network. *Med Care*, Jun 2013.
[24] J. K. Kjekshus. Importance of heart rate in determining beta-blocker efficacy in acute and long-term acute myocardial infarction intervention trials. *Am. J. Cardiol.*, 57(12):43F–49F, Apr 1986.
[25] O. Kucuktunc and et al. A large-scale sentiment analysis for Yahoo! Answers. In *Proceedings of the fifth ACM international conference on Web search and data mining*, WSDM '12, pages 633–642, New York, NY, USA, 2012. ACM.
[26] A. Leblanc and et. al. Translating comparative effectiveness of depression medications into practice by comparing the depression medication choice decision aid to usual care: study protocol for a randomized controlled trial. *Trials*, 14(1):127, May 2013.
[27] B. R. Luce and et al. Rethinking randomized clinical trials for comparative effectiveness research: the need for transformational change. *Ann. Intern. Med.*, 151(3):206–209, Aug 2009.
[28] I. G. Naglie and A. S. Detsky. Treatment of chronic nonvalvular atrial fibrillation in the elderly: a decision analysis. *Med Decis Making*, 12(4):239–249, 1992.
[29] D. Nguyen, N. A. Smith, and C. P. Rosé. Author age prediction from text using linear regression. In *Proceedings of the 5th ACL-HLT Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*, LaTeCH '11, pages 115–123, Stroudsburg, PA, USA, 2011. Association for Computational Linguistics.
[30] I. of Medicine Committee on Comparative Effectiveness Research Prioritization. *Initial National Priorities for Comparative Effectiveness Research*. 2009.
[31] M. J. Paul and M. Dredze. You Are What You Tweet : Analyzing Twitter for Public Health. *Artificial Intelligence*, pages 265–272, 2011.
[32] J. Pei and et al. Prefixspan: Mining sequential patterns by prefix-projected growth. In *Proceedings of the 17th International Conference on Data Engineering*, pages 215–224, Washington, DC, USA, 2001. IEEE Computer Society.
[33] A. Sadilek, H. A. Kautz, and V. Silenzio. Predicting disease transmission from geo-tagged micro-blog data. In *AAAI*, 2012.
[34] J. V. Selby, A. C. Beal, and L. Frank. The Patient-Centered Outcomes Research Institute (PCORI) national priorities for research and initial research agenda. *JAMA*, 307(15):1583–1584, Apr 2012.
[35] K. Shimamoto and M. Kawana. [The Carvedilol and ACE-Inhibitor Remodelling Mild Heart Failure Evaluation Trial]. *Nippon Rinsho*, 65 Suppl 4:531–536, Apr 2007.
[36] K. Sliwa and et al. Impact of initiating carvedilol before angiotensin-converting enzyme inhibitor therapy on cardiac function in newly diagnosed heart failure. *J. Am. Coll. Cardiol.*, 44(9):1825–1830, 2004.
[37] P. Smith, H. Arnesen, and I. Holme. The effect of warfarin on mortality and reinfarction after myocardial infarction. *N. Engl. J. Med.*, 323(3):147–152, Jul 1990.
[38] H. C. Sox and S. Greenfield. Comparative effectiveness research: a report from the Institute of Medicine. *Ann. Intern. Med.*, 151(3):203–205, Aug 2009.
[39] E. C. Stecker and et al. Prophylactic pacemaker use to allow beta-blocker therapy in patients with chronic heart failure with bradycardia. *Am. Heart J.*, 151(4):820–828, Apr 2006.
[40] S. Toh and et al. Confounding Adjustment in Comparative Effectiveness Research Conducted Within Distributed Research Networks. *Med Care*, Jun 2013.
[41] N. Wilcken, J. Hornbuckle, and D. Ghersi. Chemotherapy alone versus endocrine therapy alone for metastatic breast cancer. *Cochrane Database Syst Rev*, (2):CD002747, 2003.
[42] B. Woo and et al. Differences in fatigue by treatment methods in women with breast cancer. *Oncol Nurs Forum*, 25(5):915–920, Jun 1998.
[43] H. Zhu and et al. Automatic extracting of patient-related attributes: disease, age, gender and race. *Stud Health Technol Inform*, 180:589–593, 2012.