Aggregating Personal Health Messages for Scalable Comparative Effectiveness Research

Jason H.D. Cho*†
cho33@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 statistically 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 permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires specific permission and/or a fee. Request permissions from Permissions@acm.org.

BCB’13 September 22 - 25 2013, Washington, DC, USA
Copyright 2013 ACM 978-1-4503-2434-2/13/09 ...$15.00
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 bottom-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 Connect [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 \(e_{d,i,w} \in E_{d,i}\).

As an illustrative example, surrounding of sentence \(s_{d,1}\) with \(W = 1\) includes sentences \(s_{d,1-1}, s_{d,1},\) and \(s_{d,1+1}\). 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 \(t_{d,i} \in T\) be contained in sentence \(s_{d,i}\). Given treatment \(t_{d,i,w}\) that is contained in \(e_{d,i,w} \in E_{d,i}\), the surrounding of \(s_{d,i}\) is consistent if \(t_{d,i,w} \neq t_{d,i,0}\), \(\forall w \in [-W, W]\). Context \(c_{d,i}\) is defined as the union of all the surroundings with treatment \(t_{d,i,w}\) 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 sentiment, 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 \(a^t_u = \sum a_{c_{d,i}} \in A^t_u\).

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 \(\text{sgn}(a^t_u) > \text{sgn}(a^{t'}_u)\). There is no preference between treatment \(t\) to \(t'\) if \(\text{sgn}(a^t_u) = \text{sgn}(a^{t'}_u)\).

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.
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 numbers 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 by aggregating multiple resources.

Population effectiveness studies are defined as the following. Each user $u$ has sentiment towards a treatments $t$ or $t'$ 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 assign the same weights for all users who possess personal health messages.

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 $\alpha$ and $\beta$, 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: \((D, U), U_p, \alpha, \beta\)

Output: fsp, mfsp

\[
\text{phrases} = \{} \\
\text{for } (d, u) \in (D, U) \text{ do} \\
\quad \text{if } u_p \in U_p \text{ then} \\
\quad \quad \text{phrase} \leftarrow \text{getPhrases}(d, \text{age}(u_p)) \\
\quad \quad \text{phrases} = \text{phrases} \cup \text{phrase} \\
\quad \text{end if} \\
\quad \text{end for} \\
\text{end if} \\
\text{while fsp not changed do} \\
\quad \text{for } u \in U, d \in \text{docs}(u), s \in d \text{ do} \\
\quad \quad (\text{Age, Pattern}) \leftarrow \text{getAge}(s, \text{fsp, mfsp, dwords}) \\
\quad \quad \text{updatePatternPrecision(Pattern)} \\
\quad \text{end for} \\
\quad \text{for Pattern} \in \text{fsp} \text{ below } \beta \text{ precision do} \\
\quad \quad \text{Remove Pattern from fsp} \\
\quad \quad \text{if Pattern below } \alpha \text{ precision then} \\
\quad \quad \quad \text{add Pattern to mfsp} \\
\quad \text{end if} \\
\quad \text{end for} \\
\text{end while}
\]

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 \(\alpha\). We label frequently misclassified phrases as ‘msfp.’ Frequent patterns are removed if they are below \(\beta\). These two parameters are used to control two facets of the algorithm. Higher \(\beta\) denotes how aggressive we wish to be in removing patterns whereas higher \(\alpha\) 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.

Algorithm 2 Age Inference

Input: \((D, U), fsp, mfsp, dwords\)

Output: \((u, \text{age})\)

\[
\text{potAges} = \{} \\
\text{for } u \in U \text{ do} \\
\quad \text{potAges} = \{} \\
\quad \quad \text{for } d \in \text{docs}(u), s \in d \text{ do} \\
\quad \quad \quad (\text{age, Pattern}) \leftarrow \text{getAge}(s, fsp, mfsp, dwords) \\
\quad \quad \quad \text{potAges} = \text{potAges} \cup \text{age} \\
\quad \text{end for} \\
\quad \text{end for} \\
\text{end if} \\
\text{end for}
\]

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, anti-coagulants, inhibitors and blockers. For breast cancer, three categories of treatments were compared: chemotherapy, ra-
breast cancer 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 ($0.05 < p < 0.10$) 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 treatments.

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.
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 shows the result for younger group. We found three groups of comparison that were significant in the overall analysis become not 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.

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 (χ² = 3.42, p = 0.06). For hormonal treatment, a lower proportion of younger people expressed negative opinions than older ones did (χ² = 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 existing literature for relevant evidence. We found at least
### Table 8: Demographic comparison on Heart Disease on older population

<table>
<thead>
<tr>
<th>Disease</th>
<th>Heart Disease Treatments</th>
</tr>
</thead>
<tbody>
<tr>
<td>Treatment 1 (T1)</td>
<td>Treatment 2 (T2)</td>
</tr>
<tr>
<td>N1</td>
<td>166</td>
</tr>
<tr>
<td>N2</td>
<td>217</td>
</tr>
<tr>
<td>p(positive on T1)</td>
<td>0.35</td>
</tr>
<tr>
<td>p(positive on T2)</td>
<td>0.23</td>
</tr>
<tr>
<td>χ² (Positive)</td>
<td>5.95</td>
</tr>
<tr>
<td>Degree of Freedom</td>
<td>1</td>
</tr>
<tr>
<td>p-value (Positive)</td>
<td>&lt; 0.01</td>
</tr>
<tr>
<td>p(value on T1)</td>
<td>0.41</td>
</tr>
<tr>
<td>p(negative on T1)</td>
<td>0.62</td>
</tr>
<tr>
<td>χ² (Negative)</td>
<td>15.39</td>
</tr>
<tr>
<td>Degree of Freedom</td>
<td>1</td>
</tr>
<tr>
<td>p-value (Negative)</td>
<td>&lt; 0.01</td>
</tr>
<tr>
<td>T1 &gt; T2?</td>
<td>Yes</td>
</tr>
</tbody>
</table>

### Table 9: Demographic comparison on Heart Disease on younger population

<table>
<thead>
<tr>
<th>Disease</th>
<th>Heart Disease Treatments</th>
</tr>
</thead>
<tbody>
<tr>
<td>Treatment 1 (T1)</td>
<td>Treatment 2 (T2)</td>
</tr>
<tr>
<td>N1</td>
<td>75</td>
</tr>
<tr>
<td>N2</td>
<td>120</td>
</tr>
<tr>
<td>p(positive on T1)</td>
<td>0.39</td>
</tr>
<tr>
<td>p(positive on T2)</td>
<td>0.24</td>
</tr>
<tr>
<td>χ² (Positive)</td>
<td>5.13</td>
</tr>
<tr>
<td>Degree of Freedom</td>
<td>1</td>
</tr>
<tr>
<td>p-value (Positive)</td>
<td>&lt; 0.01</td>
</tr>
<tr>
<td>p(negative on T1)</td>
<td>0.44</td>
</tr>
<tr>
<td>p(negative on T2)</td>
<td>0.58</td>
</tr>
<tr>
<td>χ² (Negative)</td>
<td>15.39</td>
</tr>
<tr>
<td>Degree of Freedom</td>
<td>1</td>
</tr>
<tr>
<td>p-value (Negative)</td>
<td>&lt; 0.01</td>
</tr>
<tr>
<td>T1 &gt; T2?</td>
<td>Yes</td>
</tr>
</tbody>
</table>

### Table 10: Demographic comparison on Heart Disease on female

<table>
<thead>
<tr>
<th>Disease</th>
<th>Heart Disease Treatments</th>
</tr>
</thead>
<tbody>
<tr>
<td>Treatment 1 (T1)</td>
<td>Treatment 2 (T2)</td>
</tr>
<tr>
<td>N1</td>
<td>351</td>
</tr>
<tr>
<td>N2</td>
<td>469</td>
</tr>
<tr>
<td>p(positive on T1)</td>
<td>0.30</td>
</tr>
<tr>
<td>p(positive on T2)</td>
<td>0.20</td>
</tr>
<tr>
<td>χ² (Positive)</td>
<td>8.65</td>
</tr>
<tr>
<td>Degree of Freedom</td>
<td>1</td>
</tr>
<tr>
<td>p-value (Positive)</td>
<td>&lt; 0.01</td>
</tr>
<tr>
<td>p(negative on T1)</td>
<td>0.59</td>
</tr>
<tr>
<td>p(negative on T2)</td>
<td>0.59</td>
</tr>
<tr>
<td>χ² (Negative)</td>
<td>13.21</td>
</tr>
<tr>
<td>Degree of Freedom</td>
<td>1</td>
</tr>
<tr>
<td>p-value (Negative)</td>
<td>&lt; 0.01</td>
</tr>
<tr>
<td>T1 &gt; T2?</td>
<td>Yes</td>
</tr>
</tbody>
</table>

### 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 anonymous 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