Comparing online community structure of patients of chronic diseases

Peter A. Gloor* and Hanuma Teja Maddali
MIT Center for Collective Intelligence,
Cambridge, USA
Email: pgloor@mit.edu
Email: hmaddali@mit.edu
*Corresponding author

Peter A. Margolis
Cincinnati Children’s Medical Center,
Cincinnati, USA
Email: peter.margolis@cchmc.org

Abstract: In this paper, we use social media as a mirror to understand hidden characteristics of patients of chronic diseases, who are expressing themselves on Twitter, Facebook, and in online forums. We compare the social network structure and emotionality of people talking about Crohn’s disease, cystic fibrosis, and Type 1 diabetes. We find that the Crohn’s community’s contributors are most negative on social media, while patients of cystic fibrosis are surprisingly positive. We also notice more centralised network structure of Twitter retweet networks and Facebook wall post networks for cystic fibrosis and Crohn’s than for Type 1 diabetes. This might indicate the strong leadership role played by their national foundations, Crohn’s and Colitis Foundation of America (CCFA) and Cystic Fibrosis Foundation (CFF).

Keywords: chronic diseases; online social media; comparative analysis; netnography.

Reference to this paper should be made as follows: Gloor, P.A., Maddali, H.T. and Margolis, P.A. (xxxx) ’Comparing online community structure of patients of chronic diseases’, Int. J. Organisational Design and Engineering, Vol. X, No. Y, pp.000–000.

Biographical notes: Peter A. Gloor is a Research Scientist at the Center for Collective Intelligence at MIT’s Sloan School of Management where he leads a project exploring Collaborative Innovation Networks. He is also Founder and Chief Creative Officer of software company galaxyadvisors, an Honorary Professor at University of Cologne, a Lecturer at Aalto University, Helsinki, and a Distinguished Visiting Professor at P. Universidad Católica de Chile.

Hanuma Teja Maddali is a Research Assistant at the MIT Center for Collective Intelligence. He holds a Master of Science in Electrical and Computer Engineering from Georgia Tech. His thesis was on inferring social structure through RFID tracking of primates.

Peter A. Margolis is a Professor of Pediatrics and the Director of Research at the James M. Anderson Center for Health System Excellence at Cincinnati Children’s Hospital Medical Center. His work encompasses the application and
study of quality improvement methods in a broad range of areas including primary and sub-specialty care, communities and public health settings to improve the health outcomes of children, families and communities.

This paper is a revised and expanded version of a paper entitled ‘Comparing online community structure of patients of chronic diseases’ by Maddali, H.T. Gloor, P. and Margolis, P., Proceedings of the 5th International Conference on Collaborative Innovation Networks COINs15, Tokyo, Japan, 12–14 March 2015.

1 Introduction

Social media has become a major means of communication for patients of chronic diseases; to stay in touch with each other, find support and learn about novel treatments to better cope with their illness. In particular, Facebook has become a major channel for community activation and peer group support. While there is a wealth of research on the positive role of social networks for patients (Christakis and Fowler, 2009; Valente, 2010), less research looks at the role of online social media for patients of chronic diseases. While researchers have been studying online patient support networks such as patientsLikeMe (Frost and Massagli, 2008), our research focuses on how patients of chronic diseases use open social networks such as Twitter and Facebook groups. In earlier work we found that while patients of Crohn’s disease actively participated on Facebook discussions by posting comments, they were surprisingly unconnected by not ‘friending’ each other on Facebook (Gloor et al., 2011). In this project, we build on this earlier work by extending the focus comparing Facebook pages on different chronic illnesses, and also look at how patients and other stakeholders talk about the same chronic diseases on Twitter and on disease-specific online forums. In particular, we focus on Facebook groups and Tweets about ‘Crohn’s’, ‘Type 1 diabetes’ (T1D) and ‘Cystic Fibrosis’ and compare T1D and cystic fibrosis specific online forums.

2 Background

As the internet has become a mirror of the real world, we are trying to leverage the mirror effect to gain more insights about the real world by studying the mirror (Barr and Conlon, 1994). Studying aggregated expressions of large groups of people on the internet will allow us to draw conclusions about the collective opinion and awareness of these crowds. Similar to mirror neurons in the brain, which reflect emotions of other people (Gallese and Goldman, 1998), using the internet as the mirror permits us to study what these large groups of people are really thinking (Aveiro et al., 2010). Just like a microscope affords an inside view into a blood sample, looking at aggregated expressions of groups of people will tell us if they have been infected with unexpected beliefs, are suffering from collective fever, or might be carriers of unexpected ideas (Barsade and Gibson, 1998; Daassi and Favier, 2007; Gloor and Cooper, 2007).

In this project, we look at different patient groups with chronic diseases, focusing in this pilot study on three groups of patients, with Crohn’s disease, T1D and cystic fibrosis. While these people as individuals have huge differences in age, gender, ethnicity and
Comparing online community structure of patients of chronic diseases

socioeconomic status, they all share one strong unified characteristic, namely the same chronic disease. In particular, by analysing what people are saying on Twitter, Facebook, Wikipedia, and blogs, we hope to gain deep insights about what they are thinking subconsciously, about what they might have in common as patients or family members and friends of patients sharing the same chronic disease.

Crohn’s disease, also known as inflammatory bowel disease, and the closely related ulcerative colitis are a group of chronic inflammatory conditions affecting colon and small intestines. Estimates range up to one million Americans afflicted with the disease, resulting in 2014 in 51,000 deaths. While there is no known cure for Crohn’s, the goal for the individual patient is to drive the disease into remission, through a combination of diet, immunosuppressant drugs, and in extreme cases surgery leading to an ostomy. It is usually diagnosed between the ages of 15 and 30, although it might be diagnosed at all ages.

Diabetes mellitus type 1 results from the destruction of insulin producing cells in the pancreas. Its cause is still unknown, about 2 million of people in the USA are afflicted. Symptoms are weight loss, increased thirst and hunger, and dry mouth; long-term consequences are kidney failure, heart disease, stroke, and damage to the eyes. It is caused by different combinations of genes, although environmental factors can also influence its outbreak. It is managed by insulin injections and careful tracking of the insulin level in combination with the diet.

Cystic fibrosis, also known as mucoviscidosis, is a genetic condition that has no cure. It affects the epithelial cells and may clog up the lungs. While just a few years ago the survival rate of patients was at about 20 years, it is now, thanks to improved medical care, between 37 and 50 years in the developed world. It is usually diagnosed at a young age. Patients afflicted with the disease spend two to three hours a day in therapy just to cope with the disease and cough up sputum. Frequently, cystic fibrosis patients also develop diabetes.

3 Analysing Facebook pages

Facebook pages are an excellent source of ethnographic information since they are public and accessible through an API. Because they are usually the official pages of organisations, they are more likely to be moderated compared to Facebook groups, thus also containing less spam. In an in-depth search on Facebook, we found 517 Facebook pages about cystic fibrosis, 275 groups about T1D, and 587 groups about Crohn’s disease. Table 1 lists the key statistics of the top four Facebook pages for each of the three chronic diseases. Crohn’s is the most active, as it has gathered close to 10,000 messages in just seven months.

Using the number of likes as a proxy shows that the top cystic fibrosis and Crohn’s disease pages, with around 180,000 and 90,000 likes each, are doing a relatively better job of community (patient/caregiver/supporter) activation online when compared with T1D pages. Using the number of Google search hits restricted to Facebook as a rough proxy for the three diseases, ‘T1D’ has 134,000 hits on Facebook, ‘cystic fibrosis’ has 343,000 Google hits on Facebook, and ‘Crohn’s’ has 229,000, confirming the ranking through Facebook likes. Unlike for Crohn’s and cystic fibrosis, the top T1D pages are also ‘unofficial’, not associated with a particular organisation.
Table 1  Statistics of top four Facebook pages for each of Crohn’s, T1D and cystic fibrosis

<table>
<thead>
<tr>
<th>Chronic disease</th>
<th>Facebook page name</th>
<th>Number of likes</th>
<th>Actors</th>
<th>Messages</th>
<th>Data time range</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cystic fibrosis</td>
<td>Cystic Fibrosis Foundation</td>
<td>188,194</td>
<td>4,302</td>
<td>5,669</td>
<td>Sep. 23, 2011 to Nov. 28, 2014</td>
</tr>
<tr>
<td></td>
<td>Cystic Fibrosis Trust</td>
<td>64,527</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>CysticFibrosis.com</td>
<td>17,567</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Cystic Fibrosis Canada</td>
<td>9,080</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Type 1 diabetes</td>
<td>Type 1 Diabetes</td>
<td>37,643</td>
<td>3,906</td>
<td>9,217</td>
<td>Jan. 18, 2010 to Nov. 28, 2014</td>
</tr>
<tr>
<td></td>
<td>Cure Type 1 Diabetes</td>
<td>8,636</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>I hate Diabetes (type 1)</td>
<td>6,422</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Type 1 Diabetes Awareness</td>
<td>4,960</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Crohn’s</td>
<td>Crohn’s and Colitis UK</td>
<td>90,989</td>
<td>6,143</td>
<td>9,836</td>
<td>May 1, 2014 to Nov. 28, 2014</td>
</tr>
<tr>
<td></td>
<td>CCFA – Crohn’s and Colitis Foundation of America</td>
<td>89,481</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Crohn’s and Colitis Awareness</td>
<td>30,265</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Living with Crohn’s Disease: Healthline</td>
<td>17,451</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

The data from posts on the walls of the Facebook pages was used to construct a network visualisation with Condor, a social networking tool (Gloor and Zhao, 2004). Nodes in the graph represent members of Facebook. The central node for each of the clusters in Figure 1 is the Facebook page. Given a list of comments on a post, we added an edge going from a commenter to the original poster and another from a commenter to the preceding commenter if she/he is not the same person. The visualisation reveals bridge nodes between different pages. These are users who respond to content on multiple pages within a single disease community as well as across multiple disease communities.

Figure 1  Combined network structure of three patient communities of top four Facebook pages (see online version for colours)
The connecting nodes between the pages are people posting on more than one Facebook page. As Figure 1 shows, there are people posting about more than one disease, for instance both about Crohn’s and cystic fibrosis. Figure 2 lists the key structure and content-based statistics (defined in Appendix).

**Figure 2** Key network metrics of three disease groups on Facebook (see online version for colours)

The T1D Facebook pages are the most centralised and have the highest density. This suggests that a few people might dominate the discussion, also acting as bridges between the different T1D groups – this is also visible in the graph in Figure 1. The largest Crohn’s and cystic fibrosis groups are operated by official patient organisations, which might lead to wider and less centralised communication with people sticking to communicating on the same page. T1D has the most negative sentiment on Facebook, while Crohn’s is the most emotional, and cystic fibrosis posts are using the most complex language.

Analysing the Facebook pages of one representative cystic fibrosis group shows that communication on the wall is happening on two levels. When analysing the full ‘unpruned’ network, we observe a large number of people just broadcasting, i.e., like Twitter, giving a comment on a topic without getting any response. This network is shown in Figure 3, with the page shown in the centre of the star network at right as the connecting node for otherwise unconnected actors. An automated content analysis with Condor reveals that the discussion is relatively superficial, with ‘thanks’, ‘story’, and other general words dominating. Keywords are automatically calculated using the information retrieval algorithms of our Condor tool, the colour denotes positivity of the context of the word, taken as the six to eight words at left and right of the keyword.
Figure 3  All posts on the Facebook wall of a cystic fibrosis group (word cloud coloured according sentiment (Brönnimann, 2014), the more red, the more negative, the more green, the more positive is the context of the word, size of word is frequency of the word) (see online version for colours)

Notes: The network at right shows the link structure, with the page ID in the centre. All other nodes represent a person, links correspond to one person responding to the post of another person.

Figure 4 shows the connected sub-group of the same Facebook group, with only actors and their posts kept if one actor responds to another. In Twitter, this would be similar to a retweet network. We see a dramatic drop in network density, and a lot of unconnected clusters. This means that communication, at least on the Facebook group page, seems to be happening in small clusters or ‘islands’. On the other hand, the content of the discussion changes from generic words to first names, illustrating the family like character of the group of Facebook activists who know each other.

Figure 4  Same Facebook group as in Figure 3, with posts by unconnected people removed (see online version for colours)

Based on the comparison between the open all-inclusive, and the hidden connected layer, two insights stand out:

- there is a large group of people just sharing their pain (broadcasting into the ‘void’)
- there is a small sub-community consisting of clusters of not necessarily connected groups where members are very supportive of each other.
Comparing online community structure of patients of chronic diseases

Based on these observation, an obvious recommendation to the administrators of the Facebook groups, and to patient advocates of cystic fibrosis in general would be to connect these unconnected clusters, for instance, by organising idea jams or other virtual events to bring people together.

4 Analysing Twitter

Besides Facebook groups, we also analysed the Twitter network of people tweeting about these three chronic diseases. We did two separate types of analysis, collecting

1. a full one-week snapshot of all the tweets on the three chronic diseases
2. a 18 month subsample of all the tweets using a ‘Twitter gardenhose’ collecting a 1% subset of all the tweets.

We filtered for the Twitter search words ‘diabetes’, ‘T1D’, ‘cystic fibrosis’ and ‘crohn’ and manually cleaned false positives. Through repeatedly calling the Twitter search API using the software tool Condor (Gloor and Zhao, 2004) for one week we got a complete one-week sample with all the tweets containing the search words. Figure 5 illustrates the retweet network for the week Nov. 23 to Nov. 30, 2014. This week was chosen randomly to illustrate activity for one week. Each node in the network is a person, a connecting line means that one actor mentioned another in a tweet, or retweeted a tweet by the other actor. As Figure 5 shows, most tweets are made into the void, i.e., they do not trigger any reaction. Tweets about the three diseases each have a connected component, which is most dense for T1D, similar to the Facebook pages, suggesting that there must be a small cluster of T1D ‘activists’ retweeting each other. There are also some people acting as connectors between the different disease groups, tweeting about two different diseases. The most central tweeters are a mix of individual patient activists, healthcare vendors, and patient organisations.

Figure 5  Twitter network about three chronic diseases (see online version for colours)

As Figure 6 illustrates, Crohn’s patients have the most negative Twitter sentiment, and the highest emotionality on Twitter. Crohn’s tweeters are also the most responsive, i.e., they exhibit the lowest average response time (ART) and also have the most similar
tweeting contribution pattern, i.e., the lowest variance in contribution index (AWVCI). T1D tweeters need the lowest number of nudges until somebody else responds to a tweet, (Gloor et al., 2014).

**Figure 6** Key network metrics of three disease groups on Twitter (see online version for colours)

Notes: See Appendix for definition of x-axis (sentiment, emotionality, betweenness, degree) middle chart is nudges, AWVC = average weighted variance in contribution index (Gloor et al., 2008), ART = average response time.

In the second Twitter analysis, we collected a subset extracted from the Twitter public streaming access which gives us 1% of all public tweets in the period from January 17, 2013 to May 19, 2015. We used the search terms ‘cystic fibrosis’, ‘T1D’, and ‘crohn’ and did manual data cleaning to eliminate false positives.
Comparing online community structure of patients of chronic diseases

Figure 7  Key network metrics of three disease groups on Twitter, (a) y-axis are sentiment (b) complexity (c) betweenness (see online version for colours)

(a) (b) (c)

Figure 8  Temporal analysis of activity, sentiment, emotionality, and complexity in Twitter for each of the diseases (see online version for colours)
Figure 7 illustrates our results, we see that the retweets – tweets that have been read at least once and have been found worthy to be retweeted – are more positive, less complex, and more central in the general retweet network. The general directionality of sentiment in the one-week full sample, and the 16-month subsample is similar, with T1D being most positive, followed by cystic fibrosis, and Crohn’s being most negative, indicating that this seems to be a general trend in Twitter over extended periods of time. We also included general tweets about ‘diabetes’ into our 16-month sample, illustrating that a more general disease that comes for most people relatively unexpected at an advanced age leads to more negative sentiment.

Figure 8 shows the temporal evolution of activity, positivity, emotionality, and complexity of the three chronic disease discussion of the 16-month sample. We find that these metrics are more fluctuating for cystic fibrosis than for the other two disease groups. The overall activity of all three diseases with about 6.5 to 7 tweets per day is approximately in the same range for all three groups. The traffic measured through activity per day has the highest variability for Crohn’s, while it is much more steady for T1D and cystic fibrosis. This confirms the pattern of T1D and cystic fibrosis as chronic diseases ‘one is borne with’ leading to more steady communication, while Crohn’s comes out of the blue at a more advanced age, leading to outbursts of frustration when patients experience a flare.

5 Analysing a dedicated T1D community

In addition to the Facebook and Twitter analysis, we also analysed a dedicated online community for patients of T1D as well as caregivers, guardians, and supporters of people affected by the disease. Unlike in a more general forum such as Facebook, a disease specific social network provides a greater sense of community and more focused topics of conversation for its users apart from a richer dataset of user information. The content generated by users of the website includes comments and answers to a new ‘Question of the Day’ every day which remains online for 30 days. The questions can have multiple classifications such as being multiple-choice or discussion-based, directed to a certain category of users such as patients or caregivers, and being user-generated or staff generated. Also, only users who answer a question may comment on them.

The anonymised dataset consists of 1,273 questions (out of which 1152 are multiple choice and the rest are comment-based) from July 19, 2011 to January 29, 2015 along with 252,959 answers and 27,539 comments for these questions. The mean number of questions answered per person is 51.5 (Figure 9). The presence of outliers who have answered more than 800 questions indicates that some users may be inclined to answer questions just to read the associated comments. Figure 10 shows a histogram of the number of responses per questions. User submitted questions are very popular; a majority (greater than 80%) of them being in the top two quartiles of all questions indicating that the community is very responsive to its members’ questions.
We also analysed the friendship network along with an anonymised user information database. The user information database contains the information specified in Table 2.

<table>
<thead>
<tr>
<th>Table 2</th>
<th>User profile attributes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age</td>
<td></td>
</tr>
<tr>
<td>Ethnicity</td>
<td></td>
</tr>
<tr>
<td>Age at diagnosis</td>
<td></td>
</tr>
<tr>
<td>Diagnosed (true/false)</td>
<td></td>
</tr>
<tr>
<td>Guardian(true/false)</td>
<td></td>
</tr>
<tr>
<td>Supporter (true/false)</td>
<td></td>
</tr>
<tr>
<td>Gender (male/female/other)</td>
<td></td>
</tr>
<tr>
<td>Profile creation time</td>
<td></td>
</tr>
</tbody>
</table>
Ethnicity and gender are fields that may be left blank. We consider the group that left ethnicity and gender blank to be relatively privacy-conscious. The proportion of such users on disease specific websites may depend on stigma associated with the disease as well. Our sample has a sizable proportion of people without gender (unknown or blank for 9.2% of users) and ethnicity (unknown or blank for 26.5% of users) information (see Figure 11).

**Figure 11**  Distribution of users in our database by gender and ethnicity (see online version for colours)

![Figure 11](image1)

**Figure 12**  Comment network (see online version for colours)

![Figure 12](image2)

Note: Nodes represent users and edges are added between users with successive comments on a question.

As Figure 12 illustrates, a small group of people are most influential. Their centrality comes mostly from posting a lot, as degree and betweenness centrality are correlated in this case.
Comparing online community structure of patients of chronic diseases

Figure 13  Trends in the usage of some keywords from the comments (see online version for colours)

As the world cloud in Figure 14 generated from the comments on the questions illustrates, the discussion is surprisingly negative. Negativity of the words in Figure 14 comes from the words’ context, using the automated sentiment classification system in Condor (Brönnimann, 2014).

Figure 14  Word cloud generated from comments on the questions (see online version for colours)

We also investigated the effects of forum ‘friends’ on user responsiveness. The friendship network was explored using Condor graph visualisation and analysis software. The network is sparse, a large subset of the users does not have any friends (seen as an ‘asteroid belt’ in Figure 15).
In our subsequent analysis of the friendship network, the following hypotheses were tested:

H₁ The higher the number of friends of an individual $\Rightarrow$ the higher the number of questions the individual will answer.

H₂ Users comment more on questions that previously have been commented on by their friends.

H₃ The higher the number of comments on articles by an individual $\Rightarrow$ the higher the number of friends the individual has.

H₄ People with similar backgrounds answer similar questions.

Addressing $H₁$, a scatter plot of user responsiveness to multiple-choice question vs. number of friends (Figure 16) shows significant correlation between the two ($r = 0.337$, p-value < 0.001). Unlike in Facebook where a post commented on by a friend appears on
Comparing online community structure of patients of chronic diseases

For $H_2$, we constructed pairs of people who have commented on the same questions

1. In multiple choice questions (1,152 questions, and around 21,600 comments): the fraction of pairs of commenter’s being friends is 0.19.

2. Discussion-based questions (121 questions, and around 6,000 comments): The fraction of pairs of commenter’s being friends is 0.0034.

Based on this we reject $H_2$ and propose Hypothesis $H_5$: forum friends are more likely to comment on the same multiple-choice questions than on the same discussion-based questions.

Figure 17 Scatter plot of number of comments on discussion-based vs. multiple-choice questions (see online version for colours)

Note: Corresponding linear model denoted by the blue line.

Figure 18 Scatter plot of number of comments on articles vs. friend count (see online version for colours)

Note: Corresponding linear model denoted by the blue line.
Also, there is a high correlation ($r = 0.854$, p-value < 0.001) between the number of comments on multiple-choice questions and discussion-based questions (Figure 17). People who comment more probably do not see the difference between giving their opinions on multiple-choice and discussion-based questions.

For $H_3$ (people with more friends generate more article comments), the plot of number of article comments vs. number of friends shows a significant correlation ($r = 0.296$, p-value < 0.001 without considering the outlier with 800 friends) between them (Figure 18). The number of comments that a user generates appears to be independent of the number of forum friends they have.

We also analysed patterns in user behaviour based on the type of question answered. From the analysis performed for $H_3$ and $H_1$ we hypothesise that the lack of a notification mechanism on the forum plays a role in the kind of people who comment on certain questions or articles. This is motivated by Facebook where there is such a mechanism of people receiving updates on their wall related to friends’ activity, which provides an explicit stimulus for responses as follow-up comments. The system we analyse is more restricted, because comments on a question are visible only after an answer has been submitted by a user. So rather than being motivated by an interest in providing feedback to friends, we assume that people answer 22 questions on the forum more out of interest in the related topic that could stem from their ethnic background, role in the community, etc. This leads to our hypothesis $H_4$: people with similar backgrounds answer similar questions. To test this we constructed common response networks using Condor and observed clustering according to user role, ethnicity, etc. The common response network is a weighted bidirectional graph. The weight of an edge from node $A$ to node $B$ is given by equation (1).

$$w(A, B) = \frac{|A_{Q_x} \cap A_{Q_y}|}{|Q_x|}$$

where

$Q_x$  Set of questions answered by node $X$. Each question has a unique Question ID.

$A_{Q_x}$  Set of answers provided by $X$ to $Q_x$. Each answer option has a unique Answer ID.

The edge weight can also be used as a metric for similarity between $A$ and $B$ (in this case, how similar $A$ is with respect to $B$ is based on the questions they have answered). Just considering a raw count of similar answers would mean that we consider users $A$ and $B$ to be similar even if the number of questions answered by $B$ is much larger than $A$. Normalising the raw count of common answers by number of questions answered by $A$ gives us a bias in favour of the person who has answered a lower number of questions i.e., we assume that if $|Q_d| > |Q_d|$ then $w(A, B) > w(B, A)$. We assume that $A$ could be more similar to $B$ given the information we have right now. However, when there is a very large difference in the number of questions answered by two nodes $A$ and $B$, where $|Q_d| \gg |Q_d|$ then it is very likely that $|A_{Q_x} \cap A_{Q_y}|$ is almost equal to $|Q_d|$ (there is a finite set of questions) and $w(A, B)$ is close to 1. This is a stronger assumption than required. We remove these edges introduced by users answering a large number of questions by applying a threshold for above the 92nd percentile of the edge weight. We also only select users who have answered a minimum number of questions over the four-year period. The top one-third of users have answered more than 13 questions (from Figure 9),
this acts as the threshold for adding nodes to the response graph. The resulting network has around 4,000 nodes and 180,000 edges.

We observed distinct trends when it comes to the group of ‘privacy conscious’ individuals mentioned above. When users are clustered by ethnicity we visually observe clustering among people with unknown ethnicity with regards to common questions answered and responses. The clustering in the common response network indicates that privacy-conscious people seem to answer the question of the day in a similar way (Figure 19). We also visually observe a similar clustering of people with unknown and blank gender labels (Figure 20).

**Figure 19** Users in the common response network coloured by ethnicity (see online version for colours)

![Figure 19](image1)

**Figure 20** Users in the common response network coloured by gender (see online version for colours)

![Figure 20](image2)

When labelling nodes by role (diagnosed/guardian/supporter) we find relatively low clustering when compared to the supporter network for both diagnosed and guardian networks. The supporter network (Figure 21) is highly clustered and much more likely to
answer questions the same way than would guardians (Figure 23) and patients (Figure 22).

**Figure 21** Users in the common response network coloured by a boolean ‘supporter’ parameter (see online version for colours)

Notes: Green indicates a supporter. Blue and red indicates all other roles.

**Figure 22** Users in the common response network coloured by a boolean ‘diagnosed’ parameter (see online version for colours)

Notes: Green indicates a patient. Blue and red indicates all other roles.
Comparing online community structure of patients of chronic diseases

Figure 23 Users in the common response network coloured by a boolean ‘guardian’ parameter (see online version for colours)

Notes: Green indicates a guardian. Blue and red indicates all other roles.

It seems that the supporters – people with no direct link to T1D as either patient or family members – show the largest clustering, suggesting that they are drawn to answer similar questions, while guardians and patients seem to be interested in all types of questions. In a follow-up analysis, it would be interesting to identify characteristics of questions of interest to supporters.

In addition to the T1D community, we also analysed a dedicated cystic fibrosis online forum.

6 Analysing a dedicated cystic fibrosis community

As there is a high risk of cross-infection for patients of cystic fibrosis when patients come into close physical proximity, face to face interaction is usually discouraged. Social media such as online forums therefore might play a more important role for social interaction for cystic fibrosis patients. The cystic fibrosis forum dataset we analysed contains around 34,000 anonymised threads dated from 03/04/2003 to 12/18/2014. The threads themselves are grouped in a range of topics such as ‘Adults’, ‘Families’, ‘Nutrition’, and ‘Pregnancy’. Table 3 shows the attributes available for each thread.

<table>
<thead>
<tr>
<th>Table 3 CF Forum thread fields</th>
</tr>
</thead>
<tbody>
<tr>
<td>Forum topic</td>
</tr>
<tr>
<td>Thread title</td>
</tr>
<tr>
<td>Created time</td>
</tr>
<tr>
<td>Replies</td>
</tr>
<tr>
<td>Views</td>
</tr>
</tbody>
</table>
There are around 1,200 threads in the ‘pregnancy’ category and 400 in ‘nutrition’. ‘Weight gain’ seems to be among the most frequently occurring topics for the nutrition category (Figure 24), which is expected since CF blocks normal digestive function and reduces the efficiency of nutrient absorption.

Figure 24  Example word cloud for thread titles corresponding to the ‘nutrition’ category (size is word frequency)

The text analysis of the thread titles of the cystic fibrosis pregnancy forum with Condor reveals a predominantly negative sentiment associated with the topic of pregnancy (Figure 25), as getting pregnant poses a major risk for a woman with cystic fibrosis. In the forum, this topic and sometimes conflicting advice are hotly debated, although the conclusions seem to be almost exclusively negative.

Figure 25  Condor-generated most frequently occurring terms for thread titles corresponding to the ‘pregnancy’ category (see online version for colours)

Note: Coloured according sentiment, the more red, the more negative, the more green, the more positive is the context of the word, size of word is frequency.
7 Conclusions

Table 4 resumes our findings comparing the different types of online media and the three diseases. These results are preliminary and speculative, as they show positive sampling bias towards patients and family members who are more extrovert and willing to use social media. The results are thus only representative for social media savvy patients of chronic disease, however in today’s internet world this is a significant subset of the entire population. It might be skewed somewhat towards higher socioeconomic status, however in the Western World the Internet is now used in all strata of society.

In this early study, we found that stakeholders in Crohn’s, which as a disease might appear ‘out of the blue’ in the life of a patient, are more emotional and negative than patients of Cystic Fibrosis and T1D, who have the disease since birth and are focused on creating and maintaining a long-term survival ecosystem, leading to a more balanced and generally more positive outlook of life. This has been shown in happiness research in behavioural economics, where people experiencing life-changing accidents such as loss of eye sight or becoming quadriplegic, after an adaption period move back to their original levels of happiness (Brickman et al., 1978).

Table 4 Comparison of results of social media analysis between Crohn’s, T1D, and CF

<table>
<thead>
<tr>
<th></th>
<th>Crohn’s disease</th>
<th>TID</th>
<th>Cystic fibrosis</th>
</tr>
</thead>
<tbody>
<tr>
<td>Facebook popularity</td>
<td>Top page is ‘official’</td>
<td>Fewest likes</td>
<td>Top page is ‘official’</td>
</tr>
<tr>
<td></td>
<td>CCFA page</td>
<td>No ‘official’ page</td>
<td>CF foundation page</td>
</tr>
<tr>
<td></td>
<td>second most likes</td>
<td></td>
<td>Most likes</td>
</tr>
<tr>
<td>FB network structure</td>
<td>Least centralised</td>
<td>Most centralised</td>
<td>Second most centralised</td>
</tr>
<tr>
<td>(betweenness)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>FB sentiment</td>
<td>Second most positive</td>
<td>Most negative</td>
<td>Most positive</td>
</tr>
<tr>
<td>Twitter popularity</td>
<td>Most popular (3,797 actors in randomised sample)</td>
<td>Least popular (1,190 actors in randomised sample)</td>
<td>Second most popular (2,134 actors in randomised sample)</td>
</tr>
<tr>
<td>Twitter network</td>
<td>Most centralised (no significant difference to CF)</td>
<td>Least centralised</td>
<td>Most centralised</td>
</tr>
<tr>
<td>structure (betweenness)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Twitter sentiment</td>
<td>Least positive</td>
<td>Most positive</td>
<td>Second most positive</td>
</tr>
<tr>
<td>Online forums</td>
<td>Negative sentiment in community</td>
<td>Negative sentiment in community</td>
<td></td>
</tr>
</tbody>
</table>
The unofficial Facebook pages on T1D in particular seem to be an outlet for patients to vent their frustration. The T1D Facebook discussion is the most negative of the three diseases, while showing the least traffic and number of posts. This is confirmed by the negativity we found in the subsequent content analysis of the T1D disease specific online forum. T1D bloggers and twitterers on the other hand are more positive. This is however a small activist group, who to a large extent communicates among themselves. One recommendation for the T1D community might thus be to recruit the T1D bloggers and twitterers to also contribute to the more informal discussions on Facebook.

The cystic fibrosis community shows the opposite structure. While the twitter network shows strong central leadership – mostly through the Cystic Fibrosis Foundation – the Facebook network shows a more democratic structure with active participation of many different patients and family members.

The Crohn’s online community shows yet another pattern, dominated by the frustration of newly diagnosed patients who tweet very negatively, complaining about their disease and whenever they experience a flare. On Facebook, the CCFA page and related Crohn’s pages show a pattern of broad participation, with patients drawn there through tweets and blog links. As their sentiment is mostly negative, one recommendation might be to post more positive stories on these channels.

In this exploratory project comparing the structure and emotionality of three chronic disease patient communities on social media, we find substantial differences between the three communities. Our results suggest valuable insights, leading to recommendations for patient activists, caregivers, and providers, on how to further leverage social media to improve the lives of patients of chronic diseases. For instance, the low centrality of the T1D network might suggest a role for a strong central patient advocacy organisation similar to the CF foundation or CCFA. As has been repeatedly shown in social network research (Leung et al., 2013), closing transitive triads and increasing the clustering coefficient is a reliable predictor of increased community feeling and happiness. The conclusion therefore is: connect the patients by providing strong leadership.

Acknowledgements

We are grateful for the support of Dana Ball, George Serbedzija, Brian Becker, and Danielle Gianferante from Unitio, Inc. and T1D Exchange.

References


Comparing online community structure of patients of chronic diseases


Appendix

1 Centrality definitions

a *Degree centrality* is simply a count of the number of links or connections an individual has to others in the network. A degree centrality value can range from 0 to n, where 0 means an actor has no connections to others in the network and is referred to as an isolate, and n is some positive number.

b *Betweenness centrality* measures the degree to which a person is on the shortest path between all actors in the network. The higher the number, the greater the number of actors who pass through that person to get to others in the network, or the more that person lies between others in the network. A betweenness centrality value can range from 0 to n, where 0 means an actor is not between anyone on the shortest path between any two actors, such is the case for actors at the end points of a star network, and to n which is a positive number.
c *Group degree centrality* values range from 0 to 1. When all actors in the network connect to only one central person the degree centrality is at its maximum value of 1, and the network looks like a star. When all actors have the same number of links to others in the network, the network looks like a circle and the group degree centrality is 0.

d *Group betweenness centrality* values also range from 0 to 1. A value of 0 indicates that all actors in the network have the same betweenness centrality as all the others in the network, which is the case in a circle graph, and 1 indicates that there is one person who sits between everyone else, as in a star graph. A line graph has a group betweenness centrality of .311.

e *Density* is a group centrality measure of the connectedness of the whole network. Density is simply the ratio of the actual number of links in the network over the total possible number of links in the network. The total number of links in a network is calculated as \( n \times (n - 1) \), where \( n \) is the number of nodes in network. This means that a node’s self-links are not counted and are excluded in calculating density. In matrix terms this means the diagonal is excluded (links 1 to 1, 2 to 2, etc.) If everyone is connected to everyone else in the network, the network has the maximum density of one. Density also varies from 0 to 1, where 0 indicates that there are no network connections and everyone is an isolate, and where 1 means that everyone in the network is connected to everyone else.

2 Content measure definitions

a *Sentiment* is a measure with a value between 0 and 1, where a 0 indicates that the message is very negative and where a 1 indicates the message is very positive (Brönnimann 2014), 0.5 is neutral. Sentiment is calculated as the average score for the whole text/tweet. Our algorithm achieves 80% accuracy.

b *Emotionality* is calculated on individual text segments and measures the degree of emotion expressed in the segment defined as negative and positive deviation from neutral sentiment. It is an important measure that elaborates sentiment. Sentiment can be neutral, because the positive and negative text segments can balance each other out. However, there can still be a high degree of emotionality being expressed in the text.

c *Complexity* looks at the probability or rarity of a term or word, or the likelihood that a single word will occur in a corpus of text. The more rare words there are in a text, the higher the complexity score.