Comparing Health Forums: User Engagement, Salient Entities, Medical Detail

Anna Guimarães
Erisa Terolli
Gerhard Weikum
aguimara@mpi-inf.mpg.de
eterolli@mpi-inf.mpg.de
weikum@mpi-inf.mpg.de
Max Planck Institute for Informatics
Saarbruecken, Germany

ABSTRACT

Health discussion forums provide valuable information about diseases, symptoms, treatments and risk factors from a patient perspective, and allow exchanging experiences and mutual support. This paper analyzes and compares three different forums, Health Boards, Patient and Reddit, using three major conditions as representative samples: blood pressure, depression and diabetes. The analysis investigates three principal dimensions: the intensity of user engagement, the explicit coverage of salient entities such as frequent symptoms or risk factors, and the degree of medical detail expressed by specific drugs and their dosages. We report on commonalities across the three forums and on key findings about how they differ in their contents and interactions.

CCS CONCEPTS

• **Applied computing** → **Health informatics**; • **Information systems** → **Social networks**.

KEYWORDS

Social Computing, Health informatics, Online health forums, Social Networks, Text tagging

ACM Reference Format:
Anna Guimarães, Erisa Terolli, and Gerhard Weikum. 2021. Comparing Health Forums: User Engagement, Salient Entities, Medical Detail. In *Companion Publication of the 2021 Conference on Computer Supported Cooperative Work and Social Computing (CSCW '21 Companion), October 23–27, 2021, Virtual Event, USA.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3462204.3481748

1 INTRODUCTION

Health discussion forums allow patients and caregivers to seek information and share experiences on medical conditions. They are often a starting point for medical questions by patients interested in checking symptoms and risk factors, and wishing to learn from others who have gone through similar experiences. This information exchange and mutual support is especially relevant for patients with chronic illness and possible complications, and for conditions that involve lifestyle changes. In addition, medical professionals occasionally join the discussion to provide advice, but first-hand accounts of health-issue experiences are valuable for both patients and professionals [2, 10].

This work analyzes and compares three popular health communities: the US-based Health Boards (healthboards.com), the UK-based Patient (patient.info) and specific subreddits (e.g., reddit.com/r/diabetes/). The former two are forums that are exclusively focused on health topics, whereas subreddits are specialized communities within the Reddit social media platform and therefore are typically more diverse in coverage, with personal support being an important component.

Our goal is to contrast the three forums on the principal dimensions of user engagement, salient entities like symptoms or risk factors, and medical detail about specific drugs and their dosages. For instance, a possible hypothesis is that subreddit discussions are more about personal stories whereas the dedicated forums go deeper into medical issues such as specific drugs and their side-effects. The paper centers on the following research questions:

- RQ1: What is the intensity of engagement from users in each community?
- RQ2: What are the salient entities, like symptoms, treatments, side-effects and risk factors, reported in the three forums, and are there significant differences between forums in some of these aspects?
- RQ3: When discussing treatments, to what extent are specific drugs and drug dosages covered in each community?

Our analysis is based on three representative samples of widespread and intensively covered conditions: high blood pressure (hypertension), depression and diabetes. These are chosen as they involve both treatment with medical drugs and concerns about lifestyle issues (both as risk factors and as effects).

2 RELATED WORK

There is ample prior work on analyzing and utilizing online content about patients, but the focus is mostly on scientific publications (Pubmed etc.) or clinical records (see, e.g., [7] and references given there).

CSCW '21 Companion, October 23–27, 2021, Virtual Event, USA

ACM ISBN 978-1-4503-8479-7/21/10.
https://doi.org/10.1145/3462204.3481748

Table 1: Total number of threads, posts and users, and average of posts per thread and users per thread for each dataset.

Source	Blood Pressure			Depression			Diabetes		
	#Threads	#Posts	#Users	#Threads	#Posts	#Users	#Threads	#Posts	#Users
Health Boards	4144	26280	3545	6650	46243	7992	2383	12392	2548
Patient	910	8502	66	6243	72689	6849	545	4440	682
r/Hypertension	482	1978	504	-	-	-	-	-	-
r/BloodPressure	720	3033	789	-	-	-	-	-	-
r/Depression	-	-	-	709116	2300273	378626	-	-	-
r/Diabetes	-	-	-	-	-	-	57216	622385	32358

Health forums have received less attention; prior work includes examining the role of caregiver support [6], querying and QA for effective search [12, 15], and the spread of misinformation [1, 5]. The influence of cultural background on how patients express themselves in Talklife, regarding medical vs. lay-user terminology, is investigated by [13].

Specialized health forums, such as TuDiabetes, were studied by [9, 11]. An example of studying specialized communities of cancer patients is the work by [8].

General-purpose social-media platforms also host health discussions. For example, [3] leverages data from Facebook and Reddit for cancer patients to create language models for user posts. The role of user engagement in discussions about schizophrenia on Twitter was studied by [4]. Another health condition that received attention is depression, for which prior work aimed to detect early indicators of potential self-harm and harm prevention [16, 17].

3 DATA COLLECTION

For this comparative study, we collected discussion threads from three kinds of forums:

1. Health Boards, a large community with message boards for 200+ different topics (healthboards.com/boards),
2. Patient, a UK-based forum covering several topics, from specific diseases to general wellness (patient.info/health),
3. Subreddits focused on the topics of blood pressure (hypertension), depression and diabetes: reddit.com/r/BloodPressure and reddit.com/r/Hypertension, reddit.com/r/Depression, reddit.com/r/Diabetes,[1].

We collected all publicly available posts up to 4 April 2020, using a web scraper for the first two forums and Reddit API querying for the four subreddits. Statistics on these datasets are given in Table 1. The whole corpus is preprocessed using efficient NLP tools to detect medical entities as described in 4.2. Annotated data is available at people.mpi-inf.mpg.de/~aguimara/health. To respect data ownership, we omit post content and user information, and instead provide the URLs of the original posts.

4 ANALYSIS AND COMPARISON OF HEALTH FORUMS

In this section, we examine the key characteristics of discussions in terms of user engagement, salient topics and medical detail.

[1]All health forums last accessed on July 21, 2021.

When comparing observed frequencies between different forums or different conditions, we ensure statistical significance by a Chi-Squared test reporting the chi-squared value and p-value. When comparing the mean values of different observations, we employ a one-way Anova test reporting the F-test statistic and p-value. Each Anova test is followed by a Games-Howell post-hoc test to show differences between pairs of observations.

4.1 RQ1: What is the intensity of engagement from users in each community?

As a first measure of engagement, we compare the *lengths of discussion threads* in each community, by the number of replies per initial post. For all three conditions, threads on Patient are significantly longer than on Health Boards and Reddit ($Hypertension \rightarrow F(2, 5533) = 1214.8, p < 0.05$; $Depression \rightarrow F(2, 60141) = 1020.1$, $p < 0.05$ $Diabetes \rightarrow F(2, 60141) = 1020.1444$, $p < 0.05$;). Reddit threads are the shortest on Hypertension and Depression; only the diabetes subreddit has longer threads than Health Boards. This suggests that despite the much larger post volume in Reddit, there is a major point for dedicated health forums where users engage in more intensive exchange of experiences.

A similar observation holds for the frequency of initial posts that receive *no replies* at all. Around 35% of submissions to the health subreddits under consideration received no replies. However, this should not be overinterpreted, as even specialized subreddit communities exhibit high user fluctuation, wide topical diversity and possible digression.

The third measure is the number of *distinct users* who participate in a thread. In this regard, subreddits show the largest numbers, and the Patient forum shows the lowest ($Hypertension \rightarrow F(2, 5533) = 1485.4$, $p < 0.05$; $Depression \rightarrow F(2, 722006) = 10106.3$, $p < 0.05$; $Diabetes \rightarrow F(2, 60141) = 1660.6$, $p < 0.05$). This indicates that Patient users are more likely to repeatedly contribute to a discussion, whereas Reddit users often give merely a single reply.

4.1.1 Key findings. Overall, despite the higher total activity on Reddit, the two dedicated health forums show higher intensity of user engagement. Threads on Health Boards and Patient are longer, more likely to get at least one reply, and users are more likely to participate several times in the same discussion.

Table 2: Frequent entities and entity categories.

Category	Blood Pressure	Diabetes	Depression
Symptoms	Headaches	Weight Change, Fatigue	Anxiety, Insomnia, Fatigue, Suicidal Thoughts
Risk Factors	Smoking, Salt, Alcohol, Stress	Obesity, Cholesterol, Family History	Alcohol, Anxiety, Stress
Complications	Heart Problems, Stroke	Eye Damage, Foot Damage	Anxiety, Panic Disorder, Suicidal Thoughts
Treatments (lifestyle)	Eating, Diet, Exercise	Eating, Exercise	Cognitive Behavioral Therapy, Exercise
Treatments (drugs)	ACE Inhibitors, Diuretics, Beta Blockers	Insulin, Metformin	SSRIs, SNRIs

Table 3: Comparing frequencies of entity categories.

Condition	Entity Group	Health Boards	Patient	Reddit	Statistical Test
Hypertension	Symptom	0.1325	0.1637	0.0208	$\chi^2(2) = 137.87,\ p < 0.05$
	Risk Factor	0.2245	0.3374	0.0715	$\chi^2(2) = 231.89,\ p < 0.05$
	Drug	0.4262	0.4549	0.0599	$\chi^2(2) = 584.19,\ p < 0.05$
	Lifestyle	0.1477	0.1901	0.0391	$\chi^2(2) = 125.40,\ p < 0.05$
	Complication	0.1663	0.2176	0.0349	$\chi^2(2) = 167.27,\ p < 0.05$
Depression	Symptom	0.2156	0.1714	0.0014	$\chi^2(2) = 99752.77,\ p < 0.05$
	Risk Factor	0.0639	0.0681	0.0032	$\chi^2(2) = 11471.34,\ p < 0.05$
	Drug	0.5072	0.2601	0.0001	$\chi^2(2) = 299327.65,\ p < 0.05$
	Lifestyle	0.0313	0.137	0.0001	$\chi^2(2) = 89432.01,\ p < 0.05$
	Complication	0.1803	0.3167	0.0047	$\chi^2(2) = 78973.72,\ p < 0.05$
Diabetes	Symptom	0.0831	0.1725	0.0022	$\chi^2(2) = 11418.62,\ p < 0.05$
	Risk Factor	0.0827	0.0606	0.0013	$\chi^2(2) = 10149.68,\ p < 0.05$
	Drug	0.4683	0.5248	0.0337	$\chi^2(2) = 16663.06,\ p < 0.05$
	Lifestyle	0.3093	0.3541	0.0154	$\chi^2(2) = 16013.34,\ p < 0.05$
	Complication	0.0869	0.1321	0	$\chi^2(2) = 112837.64,\ p < 0.05$

4.2 RQ2: What are the salient topics of each community: symptoms, risk factors, treatments, drugs etc.?

4.2.1 Entity detection method. To detect mentions of medical entities in each community, we used the method by [14], which is an efficient NLP tool for annotating biomedical text and maps mentions to UMLS (Unified Medial Language System). From the top 100 most frequent entities in each community and for each condition, we compile 5 categories of entities: symptoms, risk factors, complications, treatments related to lifestyle, and drug treatments. For this grouping, we used disease-specific pages of the Mayo Clinic Portal (mayoclinic.org/diseases-conditions) as a guideline. Typical entities for each category are shown in Table 2, and the relative frequencies of the entity categories in each forum are shown in Table 3. To understand how these categories are featured, we drill down into each of the three conditions.

4.2.2 Blood Pressure. Though the condition is often asymptomatic, headaches are a frequently mentioned symptom in all forums. Among risk factors, (high consumption of) salt is frequent in Reddit and Patient, while anxiety features strongly in Health Boards.

Not unnaturally, a good fraction of users appear to suffer from several chronic diseases, like hypertension and diabetes. Discussions with both conditions co-occurring exhibit notable differences between forums: users on Patient talk more often about nutrition than drugs, while Health Boards users focus more on drugs such as Metformin.

4.2.3 Depression. Across all three communities, depression symptoms like insomnia, fatigue and suicidal thoughts, appear in the most frequent entities. These terms can refer to symptoms, risk factors or complications alike; thus it is hard to differentiate between occurrences referring to post-diagnosis treatment or pre-diagnosis advice seeking.

Treatment plans often involve the use of antidepressants and cognitive behavioral therapy (CBT). The Patient forum contains significantly more mentions of CBT, up to twice as much as both Health Boards and Reddit.

4.2.4 Diabetes. We observed that users on Health Boards often talk about drugs, whereas Patient users have more discussion on lifestyle behavior such as nutrition and exercising.

The same trend shows up when comparing how users discuss symptoms like fatigue, thirst etc. Health Boards shows high co-occurrence frequencies of such symptoms with mentions of drugs, whereas Patient has them more correlated with terms like nutrition or exercise.

4.2.5 Key findings. Health Boards and Patient have a more clinical focus than Reddit, with much stronger coverage of treatment by drugs, across all three conditions. The focus of subreddits is mostly on symptoms (probably before diagnosis), risk factors (for complications) and also lifestyle issues (for prevention as well as treatment). This fits well with the broader themes and more diverse users of Reddit forums in general, whereas the two specialized communities appear to be centered on patients that are already under treatment by doctors. Between Health Boards and Patient, the fractions of drug coverage are similar, except for depression, where Health Boards has significantly higher values (to be revisited under RQ3).

4.3 RQ3: How much medical detail is given in each community: specific drugs and dosages?

The discussion of RQ2 showed that medical drugs are frequently mentioned, mostly in Health Boards and Patient (and to a much lower degree in the subreddits). We drilled down into which specific drugs or drug families are prevalent for each of the three conditions, and to what extent drug dosages are discussed as well.

For diabetes, unsurprisingly, Insulin and Metformin are prevalent across all forums ($HealthBoards \rightarrow F(1, 62160) = 240.242$, $p < 0.05$; $Patient \rightarrow F(2, 10920) = 12.251$, $p < 0.05$; $Reddit \rightarrow F(2, 12020) = 10267.376$, $p < 0.05$). For depression, if drugs are mentioned, they are dominated by the SSRI family (Selective Serotonin Reuptake Inhibitor) family which includes Zoloft, Prozac, Lexapro and others ($HealthBoards \rightarrow F(1, 59850) = 106.494$, $p < 0.05$; $Patient \rightarrow F(1, 56187) = 559.541$, $p < 0.05$; $Reddit \rightarrow F(2, 5672928) = 2.563$, $p < 0.05$). The family of SNRI drugs (Serotonin–Norepinephrine Reuptake Inhibitor) appears less frequently, perhaps because it is a more recently developed medication. For blood pressure, on the other hand, we see significant differences between the prevalent drugs in Health Boards versus Patient: the former strongly features Beta Blockers (e.g., Metoprolol, Acebutolol) and the latter shows more ACE (Angiotensin-converting-enzyme) Inhibitors (e.g., Zestril, Univasc) ($HealthBoards \rightarrow F(1, 4766) = 240.242$, $p < 0.05$; $Patient \rightarrow F(1, 1090) = 12.251$, $p < 0.05$; $Reddit \rightarrow F(2, 1144330) = 10267.376$, $p < 0.05$).

Additionally, we compared drug dosages across forums. To extract this information from post texts, we identified all snippets with a numerical value followed by dosage unit such as mg, mL, puffs, drops. These are mapped to the drug mention that is closest in proximity.

For blood pressure and diabetes, no substantial differences in drug dosages were found. For depression, however, while the same antidepressants are prevalent, Health Boards and Reddit feature higher dosages. For instance, the most popular drug, Lexapro, is consumed in significantly higher dosages in Health Boards ($\mu = 29.93mg$, $\sigma = 75.74mg$) than in Patient ($\mu = 17.54$, $\sigma = 30.77$) ($t = -2.55$, $p = 0.01$). The same holds for other SSRIs like Zoloft ($HealthBoards \rightarrow \mu = 85.91mg$, $\sigma = 93.21mg$; $Patient \rightarrow \mu = 78.71mg$, $\sigma = 86.33mg$) and Prozac ($HealthBoards \rightarrow \mu = 44.99mg$, $\sigma = 112.45mg$; $Patient \rightarrow \mu = 34.19mg$, $\sigma = 40.21mg$). Between Health Boards and Reddit, no significant differences were observed.

4.3.1 Key findings. Health Boards and Patient have much higher coverage of drugs than Reddit. Depression and diabetes are largely treated with the same (families of) medications. For blood pressure, however, Health Boards and Patient exhibit two different drug families: Betablockers and ACE Inhibitors, respectively. We believe this is due to differences in regulation and medical practice in the US (Health Boards) versus UK (Patient). Regarding drug dosages, a striking observation is the significantly higher values for antidepressants in Health Boards and Reddit compared to Patient, again likely due to the different geographic foci of the respective forums.

5 CONCLUSION

While there is ample work on analyzing online communities for topics like politics, discussion in health forums have received comparatively little attention. This paper is a first step to obtaining insight into the characteristics, benefits and limitations of health communities.

Among our key findings, the most notable observation is that specialized forums like Health Boards and Patient engage more on discussing medical detail like specific drugs and their dosages. In contrast, subreddits with analogous topics appear to be more diverse, with a focus on early-stage advice-seeking and mutual support. Comparing the US-based Health Boards and the UK-based Patient forum on the specific condition of depression, another key observation is that Health Boards features significantly more posts about antidepressant drugs whereas Patient devotes more attention to behavioral therapies.

Future work exploring the detailed demographics of these communities, including user attributes such as age and gender, could reveal more about their users' habits and needs. This information, combined with our initial findings, could guide the development of search and recommendation systems for patients seeking online information and support.

REFERENCES

[1] Rakesh Bal, Sayan Sinha, Swastika Dutta, Rishabh Joshi, Sayan Ghosh, and Ritam Dutt. 2020. Analysing the Extent of Misinformation in Cancer Related Tweets. In *Proceedings of the International AAAI Conference on Web and Social Media (ICWSM 2020, Vol. 14)*. AAAI Press, Virtual Event, 924–928. https://ojs.aaai.org/index.php/ICWSM/article/view/7359

[2] Munmun De Choudhury and Sushovan De. 2014. Mental Health Discourse on reddit: Self-Disclosure, Social Support, and Anonymity. *Proceedings of the International AAAI Conference on Web and Social Media* 8, 1 (May 2014). https://ojs.aaai.org/index.php/ICWSM/article/view/14526

[3] Anne Dirkson, Suzan Verberne, and Wessel Kraaij. 2019. Lexical Normalization of User-Generated Medical Text. In *Proceedings of the Fourth Social Media Mining for Health Applications (#SMM4H) Workshop & Shared Task*. Association for Computational Linguistics, Florence, Italy, 11–20. https://doi.org/10.18653/v1/W19-3202

[4] Sindhu Kiranmai Ernala, Tristan Labetoulle, Fred Bane, Michael L. Birnbaum, Asra F. Rizvi, John M. Kane, and Munmun De Choudhury. 2018. Characterizing Audience Engagement and Assessing Its Impact on Social Media Disclosures of Mental Illnesses. In *Proceedings of the International AAAI Conference on Web and Social Media Media (ICWSM 2018, Vol. 12)*. AAAI Press, Stanford, California, USA, 62–71. https://ojs.aaai.org/index.php/ICWSM/article/view/15027

[5] Amira Ghenai and Yelena Mejova. 2018. Fake Cures: User-centric Modeling of Health Misinformation in Social Media. *PACMHCI* 2, CSCW (2018), 58:1–58:20. https://doi.org/10.1145/3274327

[6] Michele P Hamm, Annabritt Chisholm, Jocelyn Shulhan, Andrea Milne, Shannon D Scott, Lisa M Given, and Lisa Hartling. 2013. Social media use among patients and caregivers: a scoping review. *BMJ Open* 3, 5 (2013). https://doi.org/10.1136/bmjopen-2013-002819 arXiv:https://bmjopen.bmj.com/content/3/5/e002819.full.pdf

[7] Bevan Koopman and Guido Zuccon. 2019. WSDM 2019 Tutorial on Health Search (HS2019): A Full-Day from Consumers to Clinicians (materials at http:github.comielabhealth-search-tutorial).. In *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining* (Melbourne VIC, Australia) *(WSDM '19)*. Association for Computing Machinery, New York, NY, USA, 838–839.

[8] Zachary Levonian, Drew Richard Erikson, Wenqi Luo, Saumik Narayanan, Sabirat Rubya, Prateek Vachher, Loren Terveen, and Svetlana Yarosh. 2020. Bridging Qualitative and Quantitative Methods for User Modeling: Tracing Cancer Patient Behavior in an Online Health Community. In *Proceedings of the International AAAI Conference on Web and Social Media (ICWSM 2020, Vol. 14)*. AAAI Press, Virtual Event, 405–416. https://ojs.aaai.org/index.php/ICWSM/article/view/7310

[9] Michelle L Litchman, Linda S Edelman, and Gary W Donaldson. 2018. Effect of Diabetes Online Community Engagement on Health Indicators: Cross-Sectional Study. *JMIR Diabetes* 3, 2 (24 Apr 2018), e8. https://doi.org/10.2196/diabetes.8603

[10] Xiaojuan Ma, Xinning Gui, Jiayue Fan, Mingqian Zhao, Yunan Chen, and Kai Zheng. 2018. Professional Medical Advice at your Fingertips: An empirical study of an online "Ask the Doctor" platform. *PACMHCI* 2, CSCW (2018), 116:1–116:22. https://doi.org/10.1145/3274385

[11] Lena Mamykina, Drashko Nakikj, and Noemie Elhadad. 2015. Collective Sensemaking in Online Health Forums. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems* (Seoul, Republic of Korea) *(CHI '15)*. Association for Computing Machinery, New York, NY, USA, 3217–3226. https://doi.org/10.1145/2702123.2702566

[12] Alicia L. Nobles, Eric C. Leas, Mark Dredze, and John W. Ayers. 2020. Examining Peer-to-Peer and Patient-Provider Interactions on a Social Media Community Facilitating Ask the Doctor Services. In *Proceedings of the International AAAI Conference on Web and Social Media (ICWSM 2020, Vol. 14)*. AAAI Press, Virtual Event, 464–475. https://ojs.aaai.org/index.php/ICWSM/article/view/7315

[13] Sachin R. Pendse, Kate Niederhoffer, and Amit Sharma. 2019. Cross-Cultural Differences in the Use of Online Mental Health Support Forums. *PACMHCI* 3, CSCW (2019), 67:1–67:29. https://doi.org/10.1145/3359169

[14] Amy Siu, Dat Ba Nguyen, and Gerhard Weikum. 2013. Fast entity recognition in biomedical text. In *Workshop on Data Mining for Healthcare at the 19th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (SIGKDD 2013, Vol. 19)*. ACM Press, New York, NY, USA. http://hdl.handle.net/11858/00-001M-0000-0015-3A4E-1

[15] Erisa Terolli, Patrick Ernst, and Gerhard Weikum. 2020. Focused Query Expansion with Entity Cores for Patient-Centric Health Search. In *The Semantic Web (ICSW 2020)*, Jeff Z. Pan, Valentina Tamma, Claudia d'Amato, Krzysztof Janowicz, Bo Fu, Axel Polleres, Oshani Seneviratne, and Lalana Kagal (Eds.). Springer International Publishing, Cham, 547–564. https://doi.org/10.1007/978-3-030-62419-4_31

[16] David Wadden, Tal August, Qisheng Li, and Tim Althoff. 2021. The Effect of Moderation on Online Mental Health Conversations. In *Proceedings of the International AAAI Conference on Web and Social Media (ICWSM 2021, Vol. 15)*. AAAI Press, Virtual Event, 751–763. https://ojs.aaai.org/index.php/ICWSM/article/view/18100

[17] Andrew Yates, Arman Cohan, and Nazli Goharian. 2017. Depression and Self-Harm Risk Assessment in Online Forums. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (EMNLP 2017)*. Association for Computational Linguistics, Copenhagen, Denmark, 2968–2978. https://doi.org/10.18653/v1/d17-1322