Investigating Comorbidity of Mental and Physical Disorders in Online Health Forums

Maryam Abdollahyan
Barts Health NHS Trust
London, United Kingdom
maryam.abdollahyan@nhs.net

Fabrizio Smeraldi
Queen Mary University of London
Mebomine Ltd and The Alan Turing Institute
London, United Kingdom

Rashmi Patel
King's College London
London, United Kingdom

Conrad Bessant
Queen Mary University of London
Mebomine Ltd and The Alan Turing Institute
London, United Kingdom

ABSTRACT

Online health forums are increasingly used by patients to share information and discuss a broad range of health conditions. In this study, we investigate the presence of comorbidity of mental and physical disorders in users of such forums. We apply natural language processing methods to identify posts where users discuss mental health problems alongside chronic physical diseases, and use data mining techniques to explore the comorbidity patterns in these posts. We compare our findings to those reported in the literature and show how the results obtained are correlated with real-life events.

CCS CONCEPTS

• **Applied computing** → **Health informatics**; • **Information systems** → *Data mining*; • **Computing methodologies** → Natural language processing.

KEYWORDS

online health forums, data mining, health informatics, comorbidity

ACM Reference Format:
Maryam Abdollahyan, Fabrizio Smeraldi, Rashmi Patel, and Conrad Bessant. 2020. Investigating Comorbidity of Mental and Physical Disorders in Online Health Forums. In *Proceedings of the 3rd International Conference on Applications of Intelligent Systems (APPIS 2020), January 7–9, 2020, Las Palmas de Gran Canaria, Spain.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3378184.3378195

1 INTRODUCTION

The prevalence of comorbidity, defined as the presence of one or more health conditions co-occurring with a primary condition, has increased in recent years. Mental disorders are often comorbid with one or more long-term physical conditions, which renders diagnosis of such disorders, especially at an early stage, challenging [8].

APPIS 2020, January 7–9, 2020, Las Palmas de Gran Canaria, Spain

ACM ISBN 978-1-4503-7630-3/20/01…$15.00
https://doi.org/10.1145/3378184.3378195

Social media is a rich source of health information shared by patients and care providers that can supplement the data available to clinicians for the purpose of examining comorbidity of mental and physical disorders. In this study, we present an approach to identifying patterns of comorbidity in social media users using data mining techniques. Studies in this area have mostly focused on the vocabulary used by patients with particular mental or physical disorders on Twitter and Facebook [4, 10]. We instead focus on discussions taking place on online health forums. We collect millions of posts from several forums covering a wide range of conditions, and use a natural language processing (NLP) tool to annotate them. We then use trend analysis to detect comorbidity patterns in these posts, and demonstrate the potential of online health communities to provide valuable insights into this topic.

2 RELATED WORK

Traditional studies in the field of mental health rely on data collected in person by health professionals or through surveys, and are commonly limited by cost to small sample sizes and short time frames. Social media provides an alternative source of data for such studies that are relatively easy and inexpensive to collect at large scale and cover long periods of time. Moreover, this type of data, unlike other types of health data, contain information on the lived experiences of patients (users post about their mental health problems in almost real time, as they experience them within the context of their everyday lives) as well as their interactions with other users (seeking advice, offering support, etc.).

Previous research has demonstrated the feasibility of using social media data to study a variety of mental disorders. For instance, in [3], users on Twitter were asked to complete a questionnaire measuring symptoms of depression in participants, and grant access to their Tweets, which were leveraged to train a classifier for predicting the onset of depression. Similarly, in [10] and [6], Facebook users completed a survey or shared their medical records with the authors and gave them access to their status updates. These data were then used to build a regression model to estimate users' degree of depression. In some studies [4, 9], users who self-reported having been diagnosed with a mental illness were identified and data from these users were analysed. While these approaches yield labelled data, conducting surveys and manually reviewing the posts to find patients are time-consuming and labour-intensive tasks that result in small sample sizes.

Table 1: Number of users and posts retrieved from selected online health forums

Online Health Forum	Users	Posts
Crohn's Forum [1]	23,338	682,008
HealthBoards [2]	408,614	4,953,515
HealthUnlocked [3]	320,829	8,261,030
Hep Forums [4]	2,453	57,149
Inspire [5]	236,428	3,672,583
Patient.info [6]	209,464	2,612,035
Psoriasis Association [7]	3,760	11,904

Here, we analyse conversations in online health forums, as opposed to Tweets and Facebook posts. In contrast to the above work, we do not rely on surveys or self-reported diagnoses. Instead, we automatically annotate the posts using an NLP tool and identify users of interest based on their activity (e.g., posting to a message thread about a specific condition). Not many studies in this area have considered multiple conditions and fewer studies have looked at comorbidity [5]. We do not focus on a particular condition and consider a number of comorbid physical and mental disorders.

3 DATA

We collected over 20 million posts, posted between 1999 and 2019 by over 1 million users, from seven online health forums that are publicly available. Three of these forums are dedicated to a specific (usually long-term) physical condition (e.g., Crohn's Forum), while the rest cover various health conditions (e.g., Healthboards). For more details on the structure of these forums see [1]. Table 1 lists the names of these forums and the number of unique users and posts that were retrieved from each forum. The time periods before April 2003 and after April 2019 were omitted from our analyses due to insufficient data.

In addition, we composed three lists of keywords: a list containing names of long-term physical conditions (e.g., diabetes), a list containing symptoms of mental illness (e.g., anxiety), and a list containing treatments for mental disorders (e.g., Prozac). Table 2 shows the keywords used in our analyses from each list.

4 METHODS

4.1 Concept Mapping

To begin, we performed data pre-processing. This included normalising the date format, Unicode strings and whitespace characters. Next, we used MetaMap [2], an NLP tool for recognising Unified Medical Language System (UMLS) concepts in text, to annotate the posts. Since MetaMap was developed mainly for processing biomedical text and not social media posts, we tuned its parameters before running it on the data. This included restricting the mappings to only one vocabulary source (Medical Subject Headings (MeSH)) and three semantic types (Disease or Syndrome, Mental or Behavioural Dysfunction and Clinical Drug).

[1] crohnsforum.com
[2] healthboards.com
[3] healthunlocked.com
[4] hepmag.com
[5] inspire.com
[6] patient.info
[7] psoriasis-association.org.uk

Table 2: Keywords used from lists of conditions, symptoms and treatments

List	Keywords
Conditions	Crohn's, diabetes, psoriasis
Symptoms	anxiety, depression, low mood, worrying
Treatments	antidepressant, Celexa, Citalopram, Fluoxetine, Paroxetine, Prozac, Seroxat, Sertraline, SSRI, Zoloft

4.2 Keyword Trend Analysis

As preliminary analysis, we performed frequency analysis of posts mentioning depression symptoms and antidepressant drugs (from our lists of symptoms and treatments, respectively) as follows: given a set of keywords W and a set of posts M_t, where t represents time period (here, month and year), we compute the relative frequency of annotated posts $m \in \mathsf{M}_t$ that contain a keyword from W:

$$\mathcal{F}_t(\mathsf{W}) = \frac{|\{m \in \mathsf{M}_t \,|\, m \cap \mathsf{W} \neq \emptyset\}|}{|\mathsf{M}_t|}. \tag{1}$$

We smooth $\mathcal{F}_t(\mathsf{W})$ using a 3-month sliding window.

4.3 Comorbidity Analysis

A basic understanding of comorbidity patterns can be achieved by examining the fraction of users who initiate threads about physical conditions and also post about mental disorders. The first step is identifying threads about physical diseases: in the case of forums that are dedicated to a single physical disease we retrieve all the threads, while in the case of multi-topic forums, we retrieve threads with a title or a subforum name that contains a keyword from our list of physical diseases D. We order these threads based on the date of their first post. For each time point t, let $\mathsf{I}_{\mathsf{D},t}$ be the set of users who initiated one such thread *for the first time* at t. Users in $\mathsf{I}_{\mathsf{D},t}$ presumably suffer from a disease in D. For each user $i \in \mathsf{I}_{\mathsf{D},t}$, we consider the set of posts $\mathsf{M}_{i,\mathsf{W}}$ that contain a keyword from W corresponding to mental illness symptoms or treatments. Let $\tau(m)$, $m \in \mathsf{M}_{i,\mathsf{W}}$ be the time stamp of a post. We compute

$$\mathcal{R}_t(\mathsf{D},\mathsf{W}) = \frac{\sum_{i \in \mathsf{I}_{\mathsf{D},t}} [\![\mathsf{M}_{i,\mathsf{W}} \neq \emptyset \wedge \left| \min_{m \in \mathsf{M}_{i,\mathsf{W}}} \tau(m) - t \right| \leq \Delta]\!]}{|\mathsf{I}_{\mathsf{D},t}|}, \tag{2}$$

where $[\![x]\!]$ is 1 if x is true and 0 otherwise. The maximum time interval Δ between thread initiation and the first mental health-related post was determined on the basis of the experimental distribution of such intervals, described in Section 5.

Note that this ratio is fairly insensitive to the confounding effects of including/excluding particular forums during the period of interest, as there are no reasons to believe that comorbidity is discussed in one of these forums and not in the others.

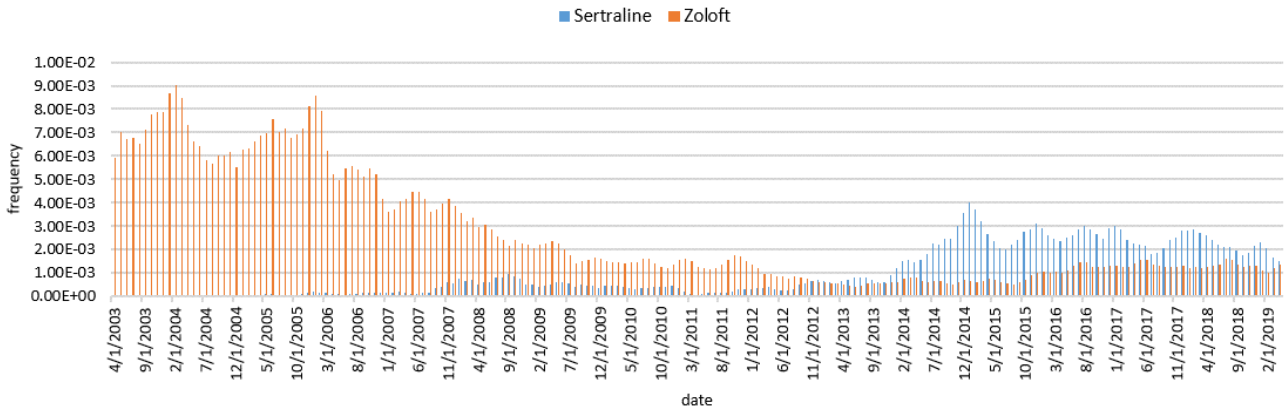

Figure 1: Frequency of posts mentioning Zoloft compared to the frequency of posts mentioning Sertraline in each month from April 2003 to April 2019

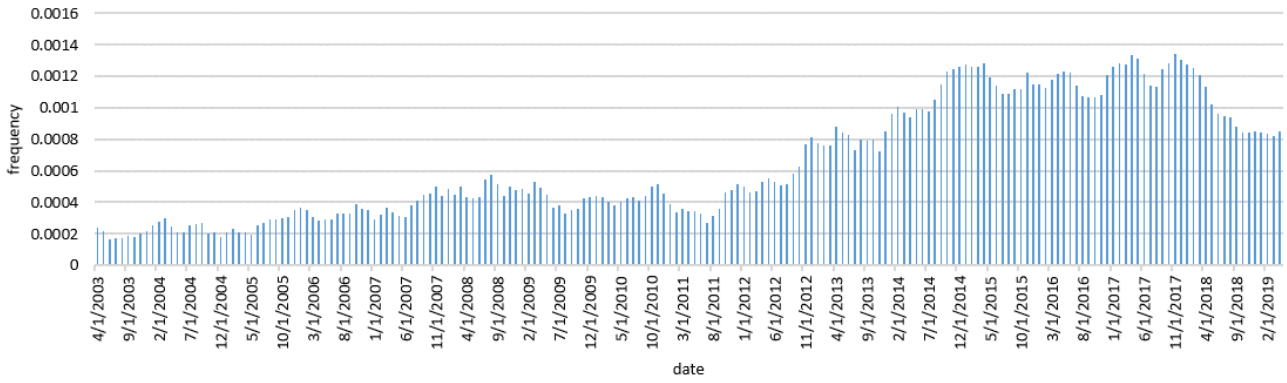

Figure 2: Frequency of posts mentioning low mood in each month from April 2003 to April 2019

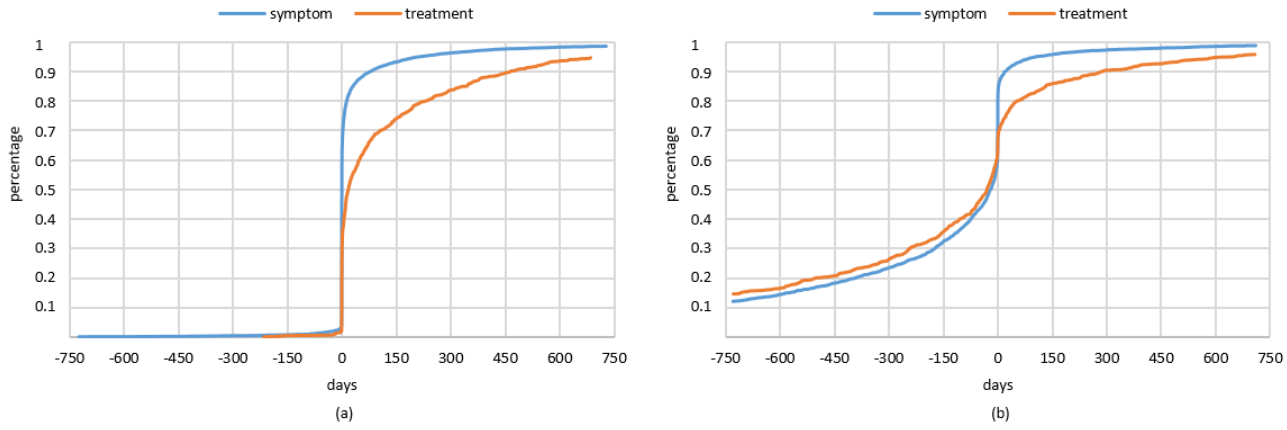

Figure 3: Cumulative distribution of the lag between users' first thread about physical diseases ($t = 0$) and their first post about mental illness symptoms or treatments in (a) Crohn's Forum and (b) HealthBoards

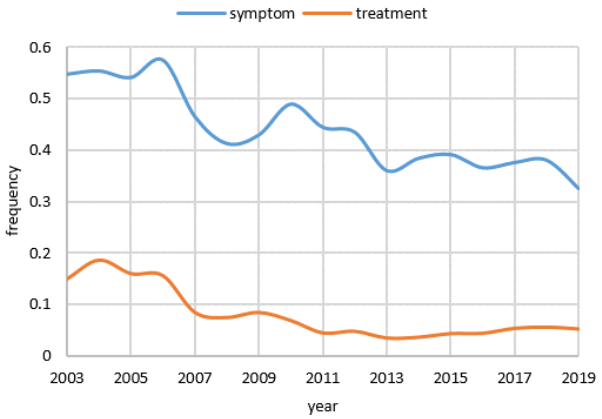

Figure 4: Fraction of physical disease thread initiators with comorbid mental disorders in each month from April 2003 to April 2019 (averaged by year)

5 RESULTS

The patterns observed in the results of keyword trend analysis correlate with real-life events. For instance, Figure 1 shows the relative frequency of posts containing the drug names Zoloft and Sertraline; a decrease in the number of mentions of Zoloft is observed after 2005, the year in which the US patent for Zoloft expired [11]; subsequently, an increase in the number of mentions of Sertraline, the corresponding generic, is observed. As another example, Figure 2 shows an increase in the relative frequency of posts mentioning 'low mood' around the time of the global financial crisis in 2008. An increase in mentions of low mood is also observed in recent years, which may reflect increased awareness and willingness to talk about mental health problems. While low mood is a symptom of depression that is comorbid with long-term physical conditions [7], a possible confounding factor here is the inclusion/exclusion of a new forum during the time period considered. This issue was addressed in Section 4.3. Interestingly, however, we did not observe an increase in the mentions of diagnoses of mental disorders (as measured by the mentions of treatments).

As for comorbidity analysis, Figures 3 (a) and 3 (b) show the time passed (in days, up to 2 years) between a user's first thread about long-term physical diseases (origin of the abscissa) and their first post about mental illness symptoms or treatments in Crohn's Forum and HealthBoards, respectively. Most users (>80%) posting about mental disorders do so within a 2-year span on either side of their first thread about physical diseases, with posts about treatments generally appearing later than the ones about symptoms. While in Crohn's Forum almost all the posts about mental health problems appear after threads about physical conditions, in HealthBoards these posts appear equally before and after such threads. This may be due to the nature of the forums: Crohn's Forum is a single-topic forum where patients initiate threads about their diagnosis, i.e., Crohn's disease, before posting about mental health problems. In contrast, HealthBoards is a multi-topic forum where patients post about various physical and mental disorders at different times.

Based on the above results, we restricted our analysis of the relative frequency of posts about comorbidity to the first post about mental health problems posted within 2 years of the first thread about a long-term physical condition, i.e., we set $\Delta = 2$ years in Equation 2. Figure 4 shows the ratio $\mathcal{R}_t(\mathrm{D}, \mathrm{W})$ for the two choices of W shown in Table 2 corresponding to mental illness symptoms and treatments. While the fraction of thread initiators who post about symptoms is higher than that of those who post about treatments, the two curves agree in terms of the overall trend. In both cases, an increase in the year following 2008 and a less pronounced increase in recent years are observed (decreases at the extremes of the time period are due to border effects).

6 CONCLUSIONS

The results of our analyses show that it is feasible to detect trends in online health forums which correlate with real-life events. The overall amplitude of these signals is quite small (absolute frequencies of posts about mental illness symptoms and treatments are of the order of a fraction of a percent), and therefore any study using this approach requires processing a large amount of data from multiple sources; hence, a solid data mining framework is needed. However, the fraction of users posting about the long-term physical conditions that we considered who also post about mental illness symptoms (>30%) or treatments (>5%) is much higher, suggesting that such forums contain a wealth of information on comorbidity of mental and physical disorders.

As indicated by the results, online health forums can offer valuable insights into the lived experiences of patients with mental health problems, particularly as users can engage in discussions without having to reveal their identity. This could facilitate discussion of comorbid mental health problems which might otherwise be avoided due to stigma. Furthermore, data from online health forums could supplement existing data from electronic case registers and observational studies to enable better understanding of individual experiences of mental illness at the population level.

ACKNOWLEDGMENTS

M. Abdollahyan was funded by the European Regional Development Fund as part of the CAP-AI project ADVancED PREPLAn (adverse event detection and patient recruitment platform). F. Smeraldi was partly supported by The Alan Turing Institute through a Fellowship.

REFERENCES

[1] Maryam Abdollahyan, Raúl J Mondragón, Conrad Bessant, and Fabrizio Smeraldi. 2018. Visualising the Topological Structure of Health-Related Message Board User Networks. In *Applications of Intelligent Systems: Proceedings of the 1st International APPIS Conference 2018*, Vol. 310. IOS Press, 274.

[2] Alan R. Aronson and François-Michel Lang. 2010. An overview of MetaMap: historical perspective and recent advances. *Journal of the American Medical Informatics Association* 17, 3 (2010), 229–236.

[3] Munmun De Choudhury, Michael Gamon, Scott Counts, and Eric Horvitz. 2013. Predicting depression via social media. In *Proceedings of the 7th International AAAI Conference on Weblogs and Social Media*.

[4] Glen Coppersmith, Mark Dredze, and Craig Harman. 2014. Quantifying mental health signals in Twitter. In *Proceedings of the Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*. 51–60.

[5] Glen Coppersmith, Mark Dredze, Craig Harman, and Kristy Hollingshead. 2015. From ADHD to SAD: Analyzing the language of mental health on Twitter through self-reported diagnoses. In *Proceedings of the 2nd Workshop on Computational*

Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality. 1–10.

[6] Johannes C. Eichstaedt, Robert J. Smith, Raina M. Merchant, Lyle H. Ungar, Patrick Crutchley, Daniel Preoţiuc-Pietro, David A. Asch, and H. Andrew Schwartz. 2018. Facebook language predicts depression in medical records. *Proceedings of the National Academy of Sciences* 115, 44 (2018), 11203–11208.

[7] Hee-Ju Kang, Seon-Young Kim, Kyung-Yeol Bae, Sung-Wan Kim, Il-Seon Shin, Jin-Sang Yoon, and Jae-Min Kim. 2015. Comorbidity of depression with physical disorders: research and clinical implications. *Chonnam Medical Journal* 51, 1 (2015), 8–18.

[8] Chris Naylor, Michael Parsonage, David McDaid, Martin Knapp, Matt Fossey, and Amy Galea. 2012. Long-term conditions and mental health: the cost of co-morbidities. (2012).

[9] Ted Pedersen. 2015. Screening Twitter users for depression and PTSD with lexical decision lists. In *Proceedings of the 2nd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*. 46–53.

[10] H Andrew Schwartz, Johannes Eichstaedt, Margaret Kern, Gregory Park, Maarten Sap, David Stillwell, Michal Kosinski, and Lyle Ungar. 2014. Towards assessing changes in degree of depression through facebook. In *Proceedings of the Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*. 118–125.

[11] Aaron Smith. 2006. Zocor and Zoloft face patent expiration. Retrieved December 12, 2019 from https://money.cnn.com/2006/06/15/news/companies/zoloft_zocor/index.htm