

Original citation:

McCaig, Duncan, Bhatia, Sudeep, Elliott, Mark T., Walasek, Lukasz and Meyer, Caroline (2018) Text-mining as a methodology to assess eating disorder-relevant factors : comparing mentions of fitness tracking technology across online communities. International Journal of Eating Disorders . doi:10.1002/eat.22882

Permanent WRAP URL:

<http://wrap.warwick.ac.uk/102764>

Copyright and reuse:

The Warwick Research Archive Portal (WRAP) makes this work by researchers of the University of Warwick available open access under the following conditions. Copyright © and all moral rights to the version of the paper presented here belong to the individual author(s) and/or other copyright owners. To the extent reasonable and practicable the material made available in WRAP has been checked for eligibility before being made available.

Copies of full items can be used for personal research or study, educational, or not-for profit purposes without prior permission or charge. Provided that the authors, title and full bibliographic details are credited, a hyperlink and/or URL is given for the original metadata page and the content is not changed in any way.

Publisher's statement:

This is the peer reviewed version of the following article: McCaig D, Bhatia S, Elliott MT, Walasek L, Meyer C. Text-mining as a methodology to assess eating disorder-relevant factors: Comparing mentions of fitness tracking technology across online communities. Int J Eat Disord. 2018. <https://doi.org/10.1002/eat.22882> , which has been published in final form at <https://doi.org/10.1002/eat.22882> . This article may be used for non-commercial purposes in accordance with [Wiley Terms and Conditions for Self-Archiving](#).

A note on versions:

The version presented here may differ from the published version or, version of record, if you wish to cite this item you are advised to consult the publisher's version. Please see the 'permanent WRAP url' above for details on accessing the published version and note that access may require a subscription.

For more information, please contact the WRAP Team at: wrap@warwick.ac.uk

1 **Title:**

2 Text-mining as a methodology to assess eating disorder-relevant factors: Comparing mentions of
3 fitness tracking technology across online communities

4

5 **Running title:**

6 Text-mining and eating disorders

7

8 **Authorship:**

9 Duncan McCaig^{a,*}, Sudeep Bhatia^b, Mark T. Elliott^a, Lukasz Walasek^a and Caroline Meyer^{a,c,d}

10 *Corresponding author. *Email address:* Duncan McCaig (d.mccaig@warwick.ac.uk)

11

12 **Authors' institutional affiliations:**

13 ^aWMG, University of Warwick, Coventry, UK

14 ^bDepartment of Psychology, University of Pennsylvania, Philadelphia, PA, USA

15 ^cWarwick Medical School, University of Warwick, Coventry, UK

16 ^dUniversity Hospitals Coventry and Warwickshire NHS Trust, UK

17

18 **Acknowledgements:**

19 Duncan McCaig is the recipient of a doctoral studentship from WMG, University of Warwick,
20 and Coventry and Warwickshire NHS Partnership Trust. Mark Elliott is currently funded by Innovate
21 UK (Project ref: 80753-504440).

1

2 **Conflict of interest statement:**

3 Mark Elliott is currently undertaking a collaborative research project with Sweatco Ltd,
4 producers of the Sweatcoin app. This project is funded by Innovate UK. Sweatco Ltd has had no
5 involvement in the research presented in this report at any stage. The remaining authors declare no
6 conflict of interest.

7

1 **Abstract**

2 *Objective:* Text-mining offers a technique to identify and extract information from a large
3 corpus of textual data. As an example, this study presents the application of text-mining to assess
4 and compare interest in fitness tracking technology across eating disorder and health-related online
5 communities. *Method:* A list of fitness tracking technology terms was developed, and communities
6 (i.e., 'subreddits') on a large online discussion platform (*Reddit*) were compared regarding the
7 frequency with which these terms occurred. The corpus used in this study comprised all comments
8 posted between May 2015 and January 2018 (inclusive) on six subreddits – three eating disorder-
9 related, and three relating to either fitness, weight-management or nutrition. All comments relating
10 to the same 'thread' (i.e., conversation) were concatenated, and formed the cases used in this study
11 ($N=377,276$). *Results:* Within the eating disorder-related subreddits, the findings indicated that a
12 'pro-eating disorder' subreddit, which is less recovery focused than the other eating disorder
13 subreddits, had the highest frequency of fitness tracker terms. Across all subreddits, the weight-
14 management subreddit had the highest frequency of the fitness tracker terms' occurrence, and
15 *MyFitnessPal* was the most frequently mentioned fitness tracker. *Discussion:* The technique
16 exemplified here can potentially be used to assess group differences to identify at-risk populations,
17 generate and explore clinically relevant research questions in populations who are difficult to
18 recruit, and scope an area for which there is little extant literature. The technique also facilitates
19 methodological triangulation of research findings obtained through more 'traditional' techniques,
20 such as surveys or interviews.

21

22 **Keywords:** eating disorders, fitness tracking, mental health, social media, text-mining

23

1 **1. Introduction**

2 Text-mining enables the identification and extraction of specific information from a large
3 corpus (i.e., collection) of textual data. This technique could provide a wealth of insight for the field
4 of eating disorders, and has previously been used in various ways within psychiatry (Abbe, Grouin,
5 Zweigenbaum, & Falissard, 2016).

6 As one of the largest online discussion platforms, *Reddit* (<https://www.reddit.com/>) offers a
7 vast source of publicly available text on which text-mining can be performed. Within *Reddit*, there
8 are ‘subreddits’ that relate to different topics (e.g., politics, films). These subreddits can be viewed
9 as communities of people with shared interests. Users can post on a subreddit to start a discussion,
10 then other people can comment on the post. This creates a series of comments in response to a post
11 (i.e., a ‘thread’), which can be conceptualised as a conversation.

12 As several subreddits relate specifically to eating disorders, the textual data available
13 through *Reddit* offers the potential to improve our understanding of these communities. Several
14 studies have qualitatively analysed the content of online eating disorder communities (e.g.,
15 Borzekowski, Schenk, Wilson, & Peebles, 2010; Branley & Covey, 2017; Sowles et al., 2018; Teufel et
16 al., 2013). This content can be broadly categorised as either: ‘pro-eating disorder’, which depicts a
17 desire to enact eating disorder behaviours without indicating a desire to recover; or ‘pro-recovery’
18 (or ‘anti-eating disorder’), which encourages recovery from eating disorders and/or resistance to
19 eating disorder cognitions and behaviours (cf. Branley & Covey, 2017). Pro-eating disorder
20 communities have been indicated to be the most active, as these groups were observed to post
21 more frequently than pro-recovery groups (Teufel et al., 2013). Regarding individual characteristics,
22 large cross-sectional surveys have found that more frequent interaction with pro-eating disorder
23 online content is associated with higher levels of eating disorder symptoms, and a higher lifetime
24 occurrence of binge eating, purging, laxative use and diet pill use (Harper, Sperry, & Thompson,
25 2008; Peebles et al., 2012). In Peebles and colleagues’ study surveying pro-eating disorder website

1 users (Peebles et al., 2012), and a separate study surveying pro-recovery website users (Aardoom,
 2 Dingemans, Boogaard, & Van Furth, 2014), both communities were found to have relatively high
 3 levels of eating disorder symptomatology, with the mean score of pro-eating disorder website users
 4 in Peebles and colleagues' study exceeding the typical clinical cut-off (e.g., Carter, Stewart, &
 5 Fairburn, 2001). Due to the differences between these online communities in terms of the extent to
 6 which they focus on recovery, the content posted in these communities can facilitate the
 7 investigation of questions regarding eating disorder readiness to change (cf. Geller, Srikameswaran,
 8 Brown, Piper, & Dunn, 2013).

9 In areas in which there is little extant research, text-mining can be particularly beneficial for
 10 scoping a research question. One such area is used to exemplify this technique – fitness tracking
 11 technology and its relationship with eating disorder symptomatology. Fitness tracking technology
 12 (i.e., fitness trackers) can be defined as devices that enable the self-monitoring of one's physical
 13 activity, such as calorie-expenditure (Simpson & Mazzeo, 2017). Recently, studies have investigated
 14 the nature of fitness tracker usage in relation to the eating disorders (Hefner et al., 2016; Levinson,
 15 Fewell, & Brosof, 2017; Simpson & Mazzeo, 2017; Tan, Kuek, Goh, Lee, & Kwok, 2016). All of these
 16 studies used self-report surveys and their findings are largely in agreement. In community samples,
 17 fitness tracker usage was found to be positively associated with eating disorder symptomatology
 18 (Hefner et al., 2016; Simpson & Mazzeo, 2017). In the context of eating disorder recovery, more
 19 clinical eating disorder patients were found to view fitness trackers as contributing to the
 20 maintenance of their eating disorder rather than aiding recovery (Tan et al., 2016). Similarly, a
 21 majority of recently discharged eating disorder patients were indicated to have used the app
 22 *MyFitnessPal*, and viewed it as having contributed to their eating disorder (Levinson et al., 2017).

23 The findings from the previously detailed survey-based studies are subject to two main
 24 limitations. First, when assessing fitness tracker usage, the majority of studies used dichotomous
 25 response scales (i.e., responding 'yes' or 'no'; Levinson et al., 2017; Simpson & Mazzeo, 2017; Tan et

1 al., 2016). Dichotomous response scales are problematic as they might increase Type I (false
 2 positive) error rate, underestimate variation in the sample, and conceal non-linearity (Altman &
 3 Royston, 2006; Austin & Brunner, 2004). These problems might cause researchers to conclude that
 4 there is a relationship when there is actually insufficient evidence. Alternatively, researchers might
 5 fail to identify a more complex relationship. Second, survey-based methodologies might have
 6 inadvertently influenced participants to give an affirmative response (e.g., “Did you feel that My
 7 Fitness Pal contributed to your eating disorder in any way?”; Levinson et al., 2017).

8 Through the application of text-mining to compare the frequency with which fitness trackers
 9 are mentioned across eating disorder subreddits, the two limitations of the previous research can be
 10 addressed. First, as frequency counts obtained through text-mining provide continuous data, no
 11 such dichotomising of measurement is undertaken. Second, as text-mining is a primarily data-driven
 12 approach, there is less risk of researchers inadvertently biasing findings. As text-mining can address
 13 these limitations, its strength is as a complementary technique, and as a way to methodologically
 14 triangulate findings obtained through more ‘traditional’ techniques (e.g., surveys). While ‘mentions’
 15 of fitness trackers cannot be assumed to reflect their usage, they can be interpreted as an indication
 16 of interest (e.g., Walasek, Bhatia, & Brown, 2017).

17 In summary, recent research concerning the relationship between eating disorder
 18 symptomatology and fitness tracker usage has two main limitations – dichotomising measurement,
 19 and potentially having influenced participants’ responses. The primary aim of the current research
 20 was to address these limitations by using text-mining to explore how pro-eating disorder and pro-
 21 recovery subreddits differed in the frequency with which fitness trackers were mentioned. A
 22 secondary aim of the research was to identify whether fitness trackers were most frequently
 23 mentioned in eating, body shape and weight, or exercise-related contexts. More generally, the
 24 current research aimed to exemplify the application of text-mining to explore an eating disorder-

1 relevant research question. As this study was exploratory in nature, no *a priori* hypotheses were
 2 stated.

3

4 **2. Methods**

5 *2.1. Overview of method*

6 A flowchart summarising the procedure is presented in Figure 1. Following selection of the
 7 corpus and generation of terms, data of interest were extracted and pre-processed, and threads
 8 were concatenated. The data were then analysed, which included summary of the corpus'
 9 characteristics, and comparison of the subreddits regarding the frequency with which different
 10 terms were mentioned.

11 [insert Fig. 1 here]

12 *2.2. Corpus selection*

13 All public *Reddit* comments since December 2005 are freely available from a regularly
 14 updated archive (Complete Public Reddit Comments Corpus, 2015). This archive does not include the
 15 initial post to which commenters responded. The entire archive between May 2015 and January
 16 2018 (inclusive) was downloaded. As fitness trackers' ubiquity and popularity vary across time, it was
 17 deemed important that all of the subreddits covered the same time period. May 2015 was selected
 18 as the start point, as this represented the earliest month at which all identified subreddits were
 19 active. Descriptions of the included subreddits are presented in Table 1. As the analyses in this study
 20 were restricted to publicly available data, an exemption from ethical review was obtained for this
 21 study from the University of Warwick's Biomedical and Scientific Research Ethics Committee.

22 *2.2.1. Eating disorder subreddits.* *Reddit's* search bar was used to identify eating disorder-related
 23 subreddits that contained one or more eating disorder-related term in the subreddit's name and/or
 24 description. The list of search terms was generated through consultation of two clinical references

1 (DSM-V; American Psychiatric Association, 2000; ICD-10; World Health Organization, 1993), and
 2 previous research (e.g., Branley & Covey, 2017; Chancellor, Lin, Goodman, Zerwas, & De Choudhury,
 3 2016). The following search terms were developed that related to eating disorders in general:
 4 “eating disorder”, “eating disorders”, “eating disordered”, “eatingdisorder”, “eatingdisorders”,
 5 “eatingdisordered”, “disordered eating”, “disordereating”, “ed”, “eds”, “proed”, “proeds”, “pro-
 6 ed” and “pro-eds”. Due to the focus on fitness trackers, the following search terms were also
 7 developed for eating disorder diagnostic categories that included criteria relating to exercise (i.e.,
 8 Anorexia Nervosa and Bulimia Nervosa): “anorexia”, “anorexic”, “anorexics”, “proanorexia”, “pro-
 9 anorexia”, “ana”, “proana”, “pro-ana”, “bulimia”, “bulimic”, “bulimics”, “probulimia”, “pro-bulimia”,
 10 “mia”, “promia” and “pro-mia”.

11 From the resultant list of eating disorder subreddits, the three with the most threads (see
 12 section 2.5.) were selected for inclusion. These were *r/proED*, *r/fuckeatingdisorders* and
 13 *r/EatingDisorders*.

14 **2.2.2. Health-related subreddits.** Three large subreddits (i.e., >150,000 subscribers on December 31
 15 2017) were also included in the corpus to explore the context in which the fitness trackers were
 16 most frequently mentioned. Three subreddits were selected that each related to one of three eating
 17 disorder-related behaviours, or behavioural outcomes – eating (i.e., *r/nutrition*), body shape and
 18 weight (i.e., *r/loseit*, a weight-management subreddit), and exercise (i.e., *r/Fitness*).

19 [insert Table 1 here]

20 **2.3. Generation of terms**

21 Complete lists of the terms detailed below are provided as Supporting Information. In line
 22 with the data pre-processing approach (section 2.4., step 1), all terms are lowercase.

23 **2.3.1. Fitness tracker terms.** A list of nouns relating to fitness trackers was developed by consulting a
 24 comprehensive website of fitness wearables (inKin Social Fitness, 2017), Google Play ‘Get Fighting

1 Fit' and 'Get Outside' health and fitness app categories (Google, 2017), and previous literature
 2 relating to fitness trackers and eating disorders (e.g., Levinson et al., 2017). Generic terms (e.g.,
 3 "fitness tracker") and other fitness trackers of theoretical interest (e.g., "cronometer") were also
 4 added.

5 Once this list had been developed, all multiword terms were added to the list of fitness
 6 tracker terms without whitespace (e.g., "fitbitsurge"). For single-word terms, each term was first
 7 entered separately into an internet search engine. In the case of multiword terms that corresponded
 8 to a brand/make (e.g., "fitbit") and model/app (e.g., "surge"), each term was also entered
 9 separately. For each separate term, if one or more of the top three search results related to the
 10 fitness tracker, it was deemed to have sufficient brand presence to be added to the list of fitness
 11 tracker terms on its own. Once the list of terms was compiled, any numeric values in the terms were
 12 removed (e.g., "couch25k" was translated into "couchk")

13 In all subsequent analyses, exact matches of the fitness tracker terms were sought.
 14 Therefore, as the terms represent nouns, singular and plural forms of each term were generated.
 15 First, the list of terms was reviewed, and irregular plurals – i.e., not created by only appending an "s"
 16 or "es" suffix to the singular form – were created on a case-by-case basis, and added as an additional
 17 term. For example, "bellabeatleaf" was pluralised to "bellabeatleaves". Two variations of each term
 18 were then generated, which represented two plural forms comprising the suffixes 's' and 'es'
 19 appended to the singular form (e.g., "fitbits"). This approach resulted in plurals that were not
 20 necessarily correct (e.g., "fitbites"). Despite this, the approach was undertaken as it is more
 21 reproducible than manually reviewing the term list and removing any ostensibly incorrect plurals. In
 22 addition, commenters might not necessarily pluralise nouns correctly (e.g., "apple watches"). As a
 23 result, the liberal approach used here also identified commenters' incorrect plurals, which were
 24 viewed as being of equal semantic importance as correct plurals.

1 The resultant list included 169 unique fitness tracker terms. As detailed above, each fitness
 2 tracker term also had two additional plural forms (i.e., +“s” and +“es”), resulting in a total of 507
 3 terms.

4 *2.3.2. Recovery, eating, body and exercise-related terms.* Three separate lists of terms were
 5 developed that related to either eating, the body or exercise. These terms were generated by
 6 compiling a list of all related terms taken from validated self-report measures. Eating terms (e.g.,
 7 “calories”) were generated from the Eating Disorder Examination Questionnaire (EDE-Q; Fairburn &
 8 Beglin, 2008) Restraint and Eating Concerns subscales, and the Shape and Weight Concerns
 9 subscales were used for body terms (e.g., “weight”). Exercise terms (e.g., “run”) were generated
 10 from the International Physical Activity Questionnaire (Craig et al., 2003) and Compulsive Exercise
 11 Test (Taranis, Touyz, & Meyer, 2011). For each identified term, several word forms were also
 12 generated (e.g., “exercise”, “exercises”, “exercised”, “exercising”). A list of five recovery terms was
 13 also generated comprising the term “recovery” and four related word forms.

14 *2.4. Data extraction and pre-processing*

15 The procedure undertaken to process the corpus is detailed in the following steps:

- 16 1) For all the comments within the corpus, the following information was extracted: month
 17 posted, author identifier, subreddit name, thread identifier, and comment text. At the point
 18 of extraction, each unique commenter was assigned an author identifier number, which was
 19 saved in place of their username. Any comments created by the automated moderation bot
 20 (“AutoModerator”) were excluded. To avoid variations in capitalisation in the corpus, each
 21 comment’s text was translated into lowercase characters. The comment text was then pre-
 22 processed by removing the following: URL links, the phrase “[deleted]” (representing a
 23 comment deleted before archiving), punctuation, numeric characters, and common English
 24 stopwords, such as personal pronouns (e.g., “my”). In each comment’s text, all occurrences

1 of multiword fitness tracker terms (section 2.3.) were also concatenated by removing
2 whitespace (e.g., “fitbit surge” was translated into “fitbitsurge”).

3 2) The text of all comments that had the same subreddit name and thread identifier was then
4 concatenated, which produced the corpus of pre-processed threads. Threads were selected
5 rather than comments, as all comments in a thread correspond to the same initial post.
6 Therefore, comments cannot be deemed to be independent, whereas threads can.

7 All data pre-processing was undertaken using the freely available natural language toolkit
8 (Bird, Klein, & Loper, 2009), and all code was written in Python programming language (Python
9 Software Foundation, 2017). The code relating to the two steps detailed above is available in an
10 online repository (<https://github.com/mccaigduncan/Text-mining-and-eating-disorders>), and can be
11 used to replicate the data extraction and pre-processing steps with the same or different subreddits.

12 *2.5. Data analysis*

13 The following steps were used to analyse the data:

- 14 1) Characteristics of the corpus and the subreddits were calculated (e.g., number of threads
15 and commenters).
- 16 2) For the eating disorder subreddits, each thread containing at least one recovery term was
17 counted, and the percentage of threads within each subreddit that referenced recovery was
18 calculated. This method was repeated for each of the eating, body and exercise lists of terms
19 for all six subreddits in the corpus.
- 20 3) The percentage of threads referencing fitness trackers was calculated for each of the six
21 subreddits using the same method as detailed in step 2). The same process was then used to
22 calculate the percentage of threads within which each individual fitness tracker term
23 occurred. After identifying the three most frequently mentioned fitness trackers across all
24 six subreddits, all terms that occurred within the corpus that related to each specific tracker

1 were grouped (e.g., “mfp” and “myfitnesspal” for *MyFitnessPal*). Using the same method as
 2 in step 2), these groups of terms were used to calculate the percentage of threads that
 3 referenced each fitness tracker.

4

5 **3. Results**

6 *3.1. Corpus characteristics*

7 For each of the six subreddits, descriptive statistics regarding the number of threads,
 8 comments and unique commenters, and the average number of comments made by each unique
 9 commenter are presented below in Table 2.

10 [insert Table 2 here]

11 Across all six subreddits, there were a total of 377,276 threads, 7,044,686 comments and
 12 508,742 unique commenters. Each unique commenter posted on an average of one out of the six
 13 subreddits ($M=1.08$, $SD=.30$; median=1, range=1:6).

14 *3.2. Mentions of recovery in eating disorder subreddits*

15 9.75% of threads within the *r/proED* subreddit mentioned recovery, compared to 43.17% of
 16 *r/fuckeatingdisorders* threads and 50.41% of *r/EatingDisorders* threads.

17 *3.3. Mentions of eating, the body and exercise in all subreddits*

18 Figure 2 presents separately the percentage of each subreddits' threads that mentioned at
 19 least one term from each of the eating, body and exercise-related lists of terms.

20 [insert Fig. 2 here]

21 Figure 2 indicates that eating was most frequently mentioned in *r/nutrition*, the body was
 22 most frequently mentioned in *r/loseit*, and exercise was most frequently mentioned in *r/Fitness*.

1 *r/loseit* was also indicated to have the second highest proportion of threads mentioning both eating
 2 and exercise-related terms.

3 *3.4. Mentions of fitness trackers in all subreddits*

4 Figure 3 presents the percentage of threads within a subreddit that contained one or more
 5 of the fitness tracker terms.

6 [insert Fig. 3 here]

7 Regarding the eating disorder subreddits, Figure 3 indicates that 5.49% of threads within the
 8 *r/proED* subreddit contained a reference to a fitness tracker, which was greater than the other
 9 eating disorder subreddits, *r/fuckeatingdisorders* and *r/EatingDisorders* (1.88% and 2.18%,
 10 respectively). Regarding the health-related subreddits, *r/loseit* had a higher percentage (30.65%)
 11 than *r/Fitness* and *r/nutrition* (5.65% and 9.05%, respectively).

12 *3.5. Frequently mentioned fitness trackers*

13 The three most frequently mentioned fitness trackers within the corpus were identified. All
 14 of the terms that related to each identified tracker and occurred in the corpus were grouped.
 15 Accordingly, the three identified fitness trackers were *MyFitnessPal*, *Fitbit* and *Heart rate monitor*.
 16 The terms relating to each fitness tracker are provided as Supporting Information, and the data for
 17 the frequency of mentions for each specific term are available in the online repository detailed in
 18 section 2.4. Figure 4 presents the percentage of threads identified in section 3.4. that included at
 19 least one reference to each aforementioned fitness tracker.

20 [insert Fig. 4 here]

21 As shown in Figure 4, among the threads in which fitness trackers were mentioned, more
 22 than 40% included mentions of *MyFitnessPal*. This was true for all six subreddits, although such
 23 mentions were particularly prevalent in *r/loseit*.

1

2 **4. Discussion**

3 This study aimed to apply text-mining to online communication to investigate an eating
4 disorder-relevant research question. The application of this technique was exemplified by analysing
5 the relative frequencies with which fitness tracker terms were mentioned within different online
6 communities (subreddits). Within the eating disorder subreddits, the fitness tracker terms were
7 most frequently mentioned in the least recovery-focused subreddit. Within health-related
8 subreddits, the highest proportion of mentions was in the weight-management subreddit. Regarding
9 specific fitness trackers, *MyFitnessPal* was the most frequently mentioned in all subreddits, and
10 occurred in 40% or more of the threads that mentioned fitness trackers. A strength of these findings
11 is that they were obtained through the unsolicited communication of over half a million users, and
12 were therefore not influenced by biases that are common in typical quantitative and qualitative
13 methodologies.

14 Regarding eating disorder subreddits, fitness trackers were more frequently mentioned in
15 *r/proED*, than in *r/fuckeatingdisorders* or *r/EatingDisorders*. Approximately ten percent of *r/proED*
16 threads mentioned recovery, compared to approximately half of the threads in *r/fuckeatingdisorders*
17 and *r/EatingDisorders*. This suggests that *r/proED* is less recovery-focused than *r/fuckeatingdisorders*
18 and *r/EatingDisorders*. Due to the more frequent mentions of fitness trackers in *r/proED*, this finding
19 suggests that, while self-reported fitness tracker usage has been positively associated with eating
20 disorder symptomatology (Hefner et al., 2016; Simpson & Mazzeo, 2017), this relationship might be
21 more nuanced. Rather, the association between fitness tracker usage and these symptoms might be
22 moderated by a person's stage of change regarding recovery (cf. Geller et al., 2013). As fitness
23 trackers were more frequently mentioned in the least recovery-focused subreddit (*r/proED*), the
24 current findings support previous survey-based research that indicated fitness tracker usage to be

1 more associated with the maintenance of eating disorders than recovery from them (Levinson et al.,
2 2017; Tan et al., 2016).

3 Within the health-related subreddits, fitness tracker terms were most frequently mentioned
4 in the weight-management subreddit (*r/loseit*), which can be interpreted as a higher interest in
5 fitness trackers in this community (e.g., Walasek et al., 2017). As *r/loseit* was indicated to be the
6 health-related subreddit with the most frequent mentions of body terms, the findings suggest that
7 fitness tracker interest is particularly high in people with a high interest in the body. This is in line
8 with previous findings that fitness tracker usage was positively associated with shape and weight
9 concerns (Simpson & Mazzeo, 2017). Undertaking exercise for weight-management (i.e., ‘weight
10 control exercise’) is a key dimension of compulsive exercise, and has been linked with shape and
11 weight concerns in both community and clinical eating disorder samples (e.g., Noetel et al., 2016;
12 Taranis et al., 2011). A positive association between overall compulsive exercise and usage of apps
13 (including fitness trackers) has also been observed (Hefner et al., 2016). Taken together, these
14 findings suggest that shape and weight concerns, weight control exercise and fitness tracker usage
15 are likely to be inter-related. Overall, fitness trackers were more frequently mentioned in the health-
16 related subreddits than in the eating disorder subreddits, which likely reflects a narrower content
17 focus of the health-related subreddits.

18 The current study also found that *MyFitnessPal* was the most mentioned fitness tracker in
19 each subreddit, which supports a research focus on the usage of this particular fitness tracker in an
20 eating disorder population (cf. Levinson et al., 2017). Additionally, commenters in *r/proED* were
21 shown to have a higher average number of comments than commenters in *r/fuckeatingdisorders*
22 and *r/EatingDisorders*, which is in line with the finding that pro-eating disorder groups posted more
23 frequently than pro-recovery groups (Teufel et al., 2013).

24 Overall, the example presented here supports the application of text-mining to complement
25 ‘traditional’ methodologies, as the current findings converged with those previously obtained

1 through survey-based measures (Hefner et al., 2016; Levinson et al., 2017; Simpson & Mazzeo, 2017;
2 Tan et al., 2016). However, both the text-mining and survey-based findings require further
3 validation, as the data obtained through these methods might not represent actual fitness tracker
4 usage.

5 There are several limitations of the text-mining approach. First, homonymy (i.e., similarly
6 spelled words with different meanings) and polysemy (i.e., one word with several meanings) are
7 problematic for text-mining (Abbe et al., 2016). For example, “apple” could refer to the fruit or the
8 brand of fitness tracker. As such, if “apple” had been included in the current research, the term
9 could have inflated frequency counts. In order to mitigate this, terms that did not identify a fitness
10 tracker in an internet search were excluded. A second general limitation is that naturally occurring
11 language might include typographic errors or variations in spelling. As such, semantically relevant
12 terms might not be identified due to these errors. This limitation was minimised in the current
13 research by generating correct and incorrect plurals. Third, the corpus used in text-mining might be
14 subject to selection bias. For example, *Reddit* was selected as it represents a large publicly available
15 source of data. However, users of the eating disorder subreddits might differ from those who visit
16 professionally managed online forums. Alternatively, a different pattern of results might have been
17 observed if a narrower time period had been selected. In order to overcome the effect of selection
18 bias, replication of the analyses using different sources of data should be conducted. Finally, the
19 generation of terms might introduce an element of subjectivity into text-mining. For example, in the
20 current research, some fitness trackers might not have been identified. This subjectivity has been
21 accounted for by ensuring that the search terms in this report were clearly detailed (enabling
22 replication and extensions), and by making relative comparisons across subreddits.

23 Regarding the specific application of text-mining exemplified above, a limitation is that the
24 included subreddits were assumed to represent people with specific characteristics (e.g., interested
25 in weight-management). However, the actual characteristics of these commenters were unknown.

1 While this limits the conclusions in the current research, this is not an issue for text-mining as a
 2 technique, as it can be used with any large corpus of text (Abbe et al., 2016). As such, if researchers
 3 had access to a corpus of text for which they had more detail (e.g., demographics of each
 4 commenter), exactly the same technique could be undertaken. Future research could assess
 5 differences between the subreddits regarding the commenters' characteristics (e.g., levels of eating
 6 disorder symptomatology).

7 In comparison to qualitative techniques, text-mining is a rapid technique for identifying and
 8 extracting salient information, and can be used with any large corpus of text. As such, it offers
 9 several potentially beneficial applications for clinically relevant research. First, text-mining can
 10 facilitate the identification of groups that are potentially at-risk for eating disorders. As exemplified
 11 in the current research, an ostensibly pro-eating disorder community mentioned fitness trackers
 12 more frequently than pro-recovery communities. A similar application of this technique could be
 13 applied to investigate group differences in eating disorder diagnostic criteria, such as purging or
 14 laxative use. Through identifying group differences, more targeted approaches could be introduced
 15 (e.g., screening, interventions). Similarly, this technique facilitates the generation and exploration of
 16 clinically relevant research questions for which there is little extant literature, or in populations who
 17 might otherwise be difficult to recruit. For example, the pro-eating disorder population might be
 18 difficult to recruit due to being characterised as non-recovery focused (Branley & Covey, 2017).

19 The application of text-mining exemplified above could be enhanced to investigate other
 20 aspects of the textual data associated with clinically relevant variables. For example, all comments
 21 that mentioned fitness trackers could be identified and extracted using the code provided. The
 22 sentiment of these comments could then be assessed using automatic techniques (e.g., Thelwall,
 23 Buckley, & Paltoglou, 2012) and compared to comments with no mentions of fitness trackers.
 24 Alternatively, 'traditional' qualitative analyses (e.g., thematic analysis) could be undertaken on these
 25 comments to explore emergent themes.

1 From a research standpoint, a strength of text-mining is that it enables the triangulation of
 2 findings from studies that use different methodologies. As previously detailed, the current
 3 application supported findings obtained through quantitative survey-based methods (e.g., Simpson
 4 & Mazzeo, 2017). By exploring the same research question with various techniques, methodology-
 5 specific limitations can be mitigated. As exemplified above, as a data-driven approach, the
 6 application of text-mining in this study was not susceptible to the limitations of the previous survey-
 7 based studies (e.g., dichotomising measurement, potentially influencing responses). Similarly, the
 8 survey-based studies are less susceptible to the limitations of this application of text-mining (e.g.,
 9 unassessed participant characteristics, homonymy/polysemy). As each methodology addresses
 10 limitations of the other, the findings obtained using both techniques will be less susceptible to biases
 11 than findings using only one. Text-mining could also be conducted in parallel with a literature review
 12 to address a research question, and generate hypotheses for confirmatory research. This would
 13 enable a more complete overview of current evidence, particularly as the relevance of extant
 14 research findings might have diminished if published a relatively long time ago.

15 In conclusion, text-mining can be used to identify illuminating patterns in an unstructured
 16 corpus of text. As exemplified in this study, the technique can be used for purposes such as
 17 triangulating findings obtained through different methodologies, scoping areas for which there is
 18 little extant research, identifying at-risk populations, and generating and exploring research
 19 questions in populations who are difficult to recruit. Due to these potentially great benefits, and as
 20 the technique is relatively rapid to undertake, it is argued that the application of text-mining is
 21 strongly warranted in the eating disorder field.

1 **References**

- 2 Aardoom, J. J., Dingemans, A. E., Boogaard, L. H., & Van Furth, E. F. (2014). Internet and patient
 3 empowerment in individuals with symptoms of an eating disorder: a cross-sectional
 4 investigation of a pro-recovery focused e-community. *Eat Behav*, *15*(3), 350-356.
 5 doi:10.1016/j.eatbeh.2014.04.003
- 6 Abbe, A., Grouin, C., Zweigenbaum, P., & Falissard, B. (2016). Text mining applications in psychiatry:
 7 a systematic literature review. *Int J Methods Psychiatr Res*, *25*(2), 86-100.
 8 doi:10.1002/mpr.1481
- 9 Altman, D. G., & Royston, P. (2006). The cost of dichotomising continuous variables. *BMJ*, *332*, 1080.
- 10 American Psychiatric Association. (2000). *Diagnostic and Statistical Manual of Mental Disorders:*
 11 *DSM-IV-TR* (4th Ed.). Washington DC: American Psychiatric Association.
- 12 Austin, P. C., & Brunner, L. J. (2004). Inflation of the type I error rate when a continuous confounding
 13 variable is categorized in logistic regression analyses. *Stat Med*, *23*(7), 1159-1178.
 14 doi:10.1002/sim.1687
- 15 Bird, S., Klein, E., & Loper, E. (2009). *Natural Language Processing with Python*: O'Reilly Media.
- 16 Borzekowski, D. L., Schenk, S., Wilson, J. L., & Peebles, R. (2010). e-Ana and e-Mia: A content analysis
 17 of pro-eating disorder Web sites. *Am J Public Health*, *100*(8), 1526-1534.
 18 doi:10.2105/AJPH.2009.172700
- 19 Branley, D. B., & Covey, J. (2017). Pro-ana versus Pro-recovery: A Content Analytic Comparison of
 20 Social Media Users' Communication about Eating Disorders on Twitter and Tumblr. *Frontiers*
 21 *in Psychology*, *8*. doi:10.3389/fpsyg.2017.01356
- 22 Carter, J. C., Stewart, D. A., & Fairburn, C. G. (2001). Eating disorder examination questionnaire:
 23 Norms for young adolescent girls. *Behaviour Research and Therapy*, *39*(5), 625-632.
- 24 Chancellor, S., Lin, Z., Goodman, E. L., Zerwas, S., & De Choudhury, M. (2016). *Quantifying and*
 25 *Predicting Mental Illness Severity in Online Pro-Eating Disorder Communities*. Paper

- 1 presented at the Proceedings of the 19th ACM Conference on Computer-Supported
 2 Cooperative Work & Social Computing - CSCW '16.
- 3 Complete Public Reddit Comments Corpus. (2015). Retrieved from
 4 https://archive.org/details/2015_reddit_comments_corpus.
- 5 Craig, C. L., Marshall, A. L., Sjostrom, M., Bauman, A. E., Booth, M. L., Ainsworth, B. E., . . . Oja, P.
 6 (2003). International physical activity questionnaire: 12-country reliability and validity. *Med*
 7 *Sci Sports Exerc*, 35(8), 1381-1395. doi:10.1249/01.MSS.0000078924.61453.FB
- 8 Fairburn, C. G., & Beglin, S. J. (2008). Eating Disorder Examination Questionnaire (EDE-Q 6.0). In C. G.
 9 Fairburn (Ed.), *Cognitive behavior therapy and eating disorders*. (pp. 309-314). New York:
 10 Guilford Press.
- 11 Geller, J., Srikameswaran, S., Brown, K. E., Piper, W., & Dunn, E. C. (2013). The Psychometric
 12 Properties of the Readiness and Motivation Questionnaire: A Symptom-Specific Measure of
 13 Readiness for Change in the Eating Disorders. *Psychological Assessment*, 25(3), 759-768.
 14 doi:10.1037/a0032539.supp
- 15 Google. (2017). Health and Fitness Apps. Retrieved from
 16 play.google.com/store/apps/category/HEALTH_AND_FITNESS?hl=en_GB.
- 17 Harper, K., Sperry, S., & Thompson, J. K. (2008). Viewership of pro-eating disorder websites:
 18 association with body image and eating disturbances. *Int J Eat Disord*, 41(1), 92-95.
 19 doi:10.1002/eat.20408
- 20 Hefner, V., Dorros, S. M., Jourdain, N., Liu, C., Tortomasi, A., Greene, M. P., . . . Alvares, C. (2016).
 21 Mobile exercising and tweeting the pounds away: The use of digital applications and
 22 microblogging and their association with disordered eating and compulsive exercise. *Cogent*
 23 *Social Sciences*, 2(1). doi:10.1080/23311886.2016.1176304
- 24 inKin Social Fitness. (2017). Wearables. Retrieved from www.inkin.com/wearables/
- 25 Levinson, C. A., Fewell, L., & Brosof, L. C. (2017). My Fitness Pal calorie tracker usage in the eating
 26 disorders. *Eat Behav*, 27, 14-16. doi:10.1016/j.eatbeh.2017.08.003

- 1 Noetel, M., Miskovic-Wheatley, J., Crosby, R. D., Hay, P., Madden, S., & Touyz, S. (2016). A clinical
 2 profile of compulsive exercise in adolescent inpatients with anorexia nervosa. *J Eat Disord, 4*,
 3 1. doi:10.1186/s40337-016-0090-6
- 4 Peebles, R., Wilson, J. L., Litt, I. F., Hardy, K. K., Lock, J. D., Mann, J. R., & Borzekowski, D. L. (2012).
 5 Disordered eating in a digital age: eating behaviors, health, and quality of life in users of
 6 websites with pro-eating disorder content. *J Med Internet Res, 14*(5), e148.
 7 doi:10.2196/jmir.2023
- 8 Python Software Foundation. (2017). Python Language Reference (Version 3.6.3.). Available at
 9 <http://www.python.org>.
- 10 Simpson, C. C., & Mazzeo, S. E. (2017). Calorie counting and fitness tracking technology: Associations
 11 with eating disorder symptomatology. *Eat Behav, 26*, 89-92.
 12 doi:10.1016/j.eatbeh.2017.02.002
- 13 Sowles, S. J., McLeary, M., Optican, A., Cahn, E., Krauss, M. J., Fitzsimmons-Craft, E. E., . . . Cavazos-
 14 Rehg, P. A. (2018). A content analysis of an online pro-eating disorder community on Reddit.
 15 *Body Image, 24*, 137-144. doi:10.1016/j.bodyim.2018.01.001
- 16 Tan, T. N., Kuek, A., Goh, S. E., Lee, E. L., & Kwok, V. (2016). Internet and smartphone application
 17 usage in eating disorders: A descriptive study in Singapore. *Asian Journal of Psychiatry, 19*,
 18 50-55. doi:10.1016/j.ajp.2015.11.007
- 19 Taranis, L., Touyz, S., & Meyer, C. (2011). Disordered eating and exercise: development and
 20 preliminary validation of the compulsive exercise test (CET). *Eur Eat Disord Rev, 19*(3), 256-
 21 268. doi:10.1002/erv.1108
- 22 Teufel, M., Hofer, E., Junne, F., Sauer, H., Zipfel, S., & Giel, K. E. (2013). A comparative analysis of
 23 anorexia nervosa groups on Facebook. *Eating and Weight Disorders-Studies on Anorexia*
 24 *Bulimia and Obesity, 18*(4), 413-420. doi:10.1007/s40519-013-0050-y
- 25 Thelwall, M., Buckley, K., & Paltoglou, G. (2012). Sentiment strength detection for the social Web.
 26 *Journal of the American Society for Information Science and Technology, 63*(1), 163-173.

- 1 Walasek, L., Bhatia, S., & Brown, G. D. A. (2017). Positional goods and the social rank hypothesis:
2 Income inequality affects online chatter about high and low status brands on Twitter.
3 *Journal of Consumer Psychology*. doi:10.1016/j.jcps.2017.08.002
- 4 World Health Organization. (1993). *The ICD-10 Classification of Mental and Behavioural Disorders:*
5 *Clinical Descriptions and Diagnostic Guidelines*. Geneva: World Health Organization.
- 6
- 7

1 **Tables**

2

3 **Table 1.** Names and descriptions of subreddits comprising the corpus

Subreddit name	Subreddit description [†]
Eating disorder subreddits	
<i>r/proED</i>	“**This is a place to discuss eating disorders, extreme/fringe eating behaviors, thinspo and recovery.”
<i>r/fuckeatingdisorders</i>	“ Eating disorders have many misconceptions, and part of that is because those who have it hide it since those who don't have it don't understand it because no one talks about it. FED is here to confront eating disorders and provide a place for anyone to ask questions.”
<i>r/EatingDisorders</i>	“## For Awareness, Information, and Questions about Recovering from EDs. We are a pro-recovery site, and only allow approved posts.”
Health-related subreddits	
<i>r/Fitness</i>	“This subreddit is for discussion of physical fitness goals and how they can be achieved.”
<i>r/loseit</i>	“A place for people of all sizes to discuss healthy and sustainable methods of weight loss. Whether you need to lose 2 lbs or 200 lbs, you are welcome here!”
<i>r/nutrition</i>	“A place to discuss eating well.”

[†]Descriptions of subreddits were extracted verbatim from reddit.com on December 31 2017

4

5

1 **Table 2.** Characteristics of the corpus

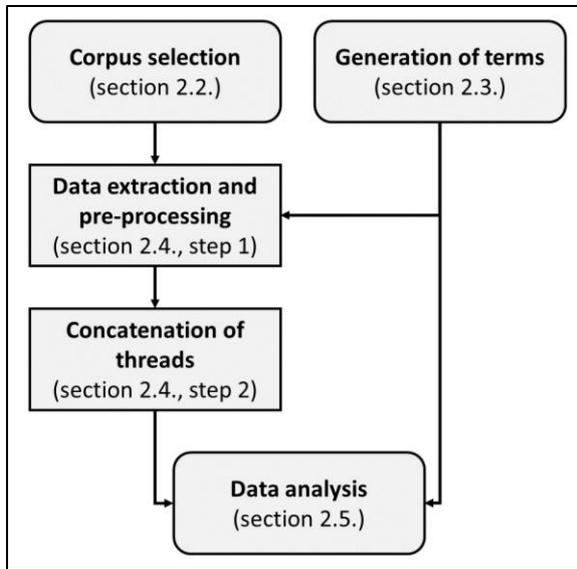
Subreddits	Number of threads	Number of comments	Number of unique commenters [†]	Average number of comments by each unique commenter [†]	
				Mean (<i>SD</i>)	Median (range)
Eating disorder					
<i>r/proED</i>	37,335	387,357	11,123	32 (104)	5 (1:2,722)
<i>r/fuckeatingdisorders</i>	1,911	10,637	1,809	6 (19)	2 (1:569)
<i>r/EatingDisorders</i>	964	4,991	1,774	3 (6)	1 (1:133)
Health-related					
<i>r/Fitness</i>	213,885	4,620,754	382,426	11 (121)	2 (1:47,835)
<i>r/loseit</i>	108,496	1,836,704	131,825	13 (123)	2 (1:21,160)
<i>r/nutrition</i>	14,685	184,243	22,847	8 (44)	2 (1:2,821)

[†]Excludes commenter '[deleted]'

2

1 **Figures**

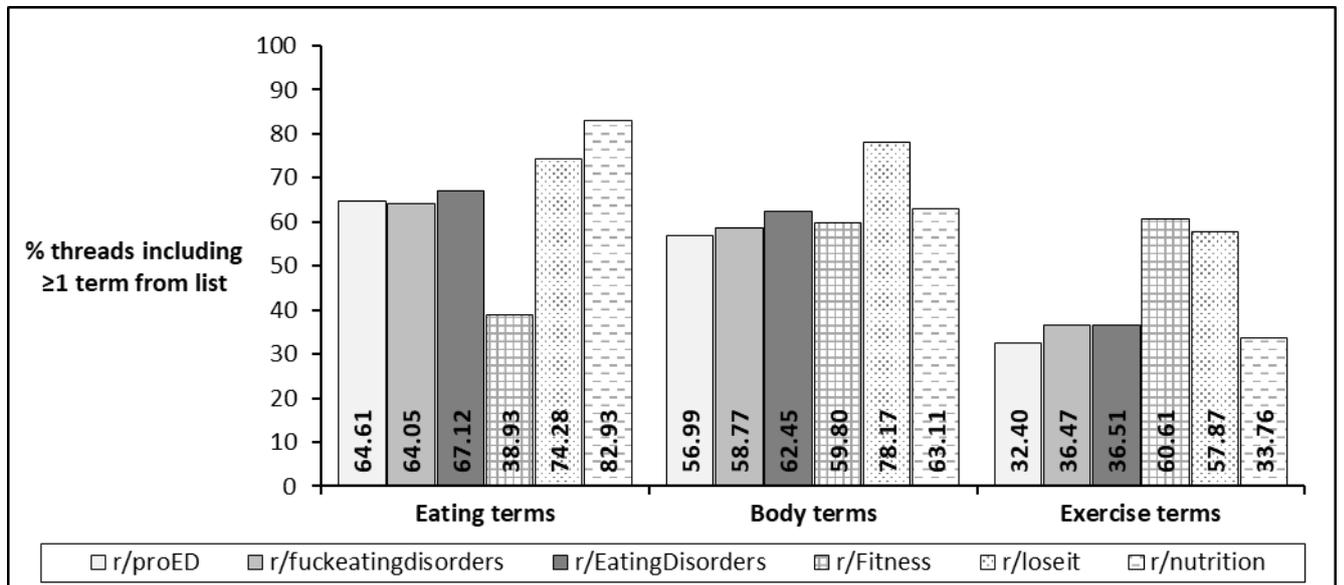
2 **Fig. 1.** Procedural flowchart



3

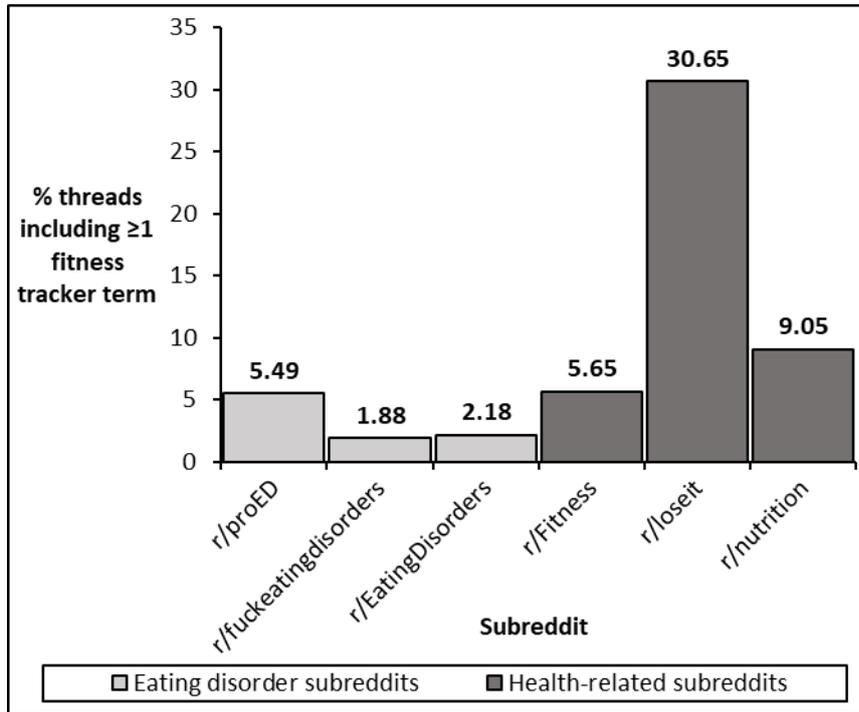
4

1 **Fig. 2.** Percentage of threads including at least one term from eating, body and exercise-related lists
 2 of terms



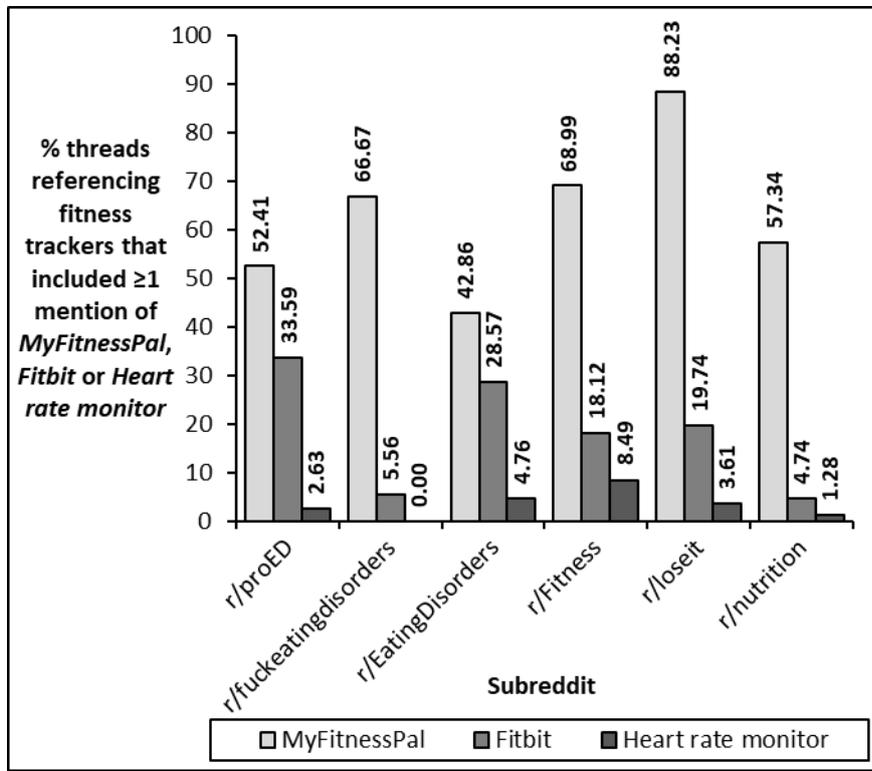
3

1 **Fig. 3.** Percentage of threads including at least one fitness tracker term



2

1 **Fig. 4.** Percentage of threads referencing fitness trackers that included at least one mention of
 2 *MyFitnessPal, Fitbit and Heart rate monitor*



3