Information Quality of Reddit Link Posts on Health News

Haichen Zhou[1,2] and Bei Yu[2(✉)]

[1] Nanjing Agricultural University, Nanjing 210095, Jiangsu, China
haizenchow@gmail.com
[2] Syracuse University, Syracuse, NY 13244, USA
byu@syr.edu

Abstract. Inaccuracy has been a common problem in news coverage of scientific research. This problem has been particularly prevalent in health research news. Health research news usually spreads from research publications and press releases to news and social media. In this study we examined the information quality of the Reddit link posts that introduce health news stories. We developed a coding schema to annotate the inaccurate information in a sample of 250 link posts on health research news within the Reddit community *r/Health* in 2018. The result shows that most link posts simply copied the original news headlines verbatim, while some paraphrased the news stories by adding, deleting, replacing, and combining content. We found that 12 paraphrased link posts contained inaccurate information that may mislead the readers. The most common type of inaccuracy is exaggeration resulted from changing the original speculative claims to direct causal statements by removing the modal verbs such as "may" and "might". The result shows that although the link posts of health news were generally faithful to the original news stories, exaggerated claims may lead to false hope for researchers and patients.

Keywords: Reddit · Health news · Paraphrasing · Information quality

1 Introduction

Inaccuracy has been a common problem in news coverage of scientific research [1, 2]. This problem has been particularly prevalent in health research news [3, 4]. Health research news usually diffuses from research publications and press releases to news and social media [3]. Inaccuracy has been attributed to journalists' lack of training or the appeal to sensationalism to arouse readers' interest [5], and has been found not only in non-credible sources like tabloids, but also in prestige newspapers and even academic press releases from the researchers' own institutions [3]. For example, exaggerated health advice, exaggerated causal claims from correlational findings and inference to humans from animal studies have been frequently found in press releases and news stories [3]. The exaggerations may result in wrong medical decisions and serious health consequences [6].

With the popularity of social media, health research news has also been spread to social media. Certain online discussion communities, such as Reddit, have become a

A. Sundqvist et al. (Eds.): iConference 2020, LNCS 12051, pp. 186–197, 2020.
https://doi.org/10.1007/978-3-030-43687-2_14

curated news source for the general public [7]. As a social news aggregation, web content rating, and discussion website[1], Reddit has a number of subreddit communities, such as *r/Health*, where participants can post links to health news stories. Different from Facebook and Twitter, Reddit implements the voting and comment functions to help the community filter out low-quality health information [8]. Reddit users are asked to provide an interesting title and text comment when post a link to a health news story[2]. They can also upvote and downvote a post. Through this collective effort, high quality content is expected to rise to the top [9]. Despite the wisdom of the crowd, many factors such as low health literacy[3] might affect Reddit voters' judgment; therefore, the accuracy of the top posts is still questionable [10–12]. Some studies have raised concern about this issue. For example, health information retrieved through social media was ranked as the least reliable source compared with information from physicians, family/friends and web search results [11]. In addition, health-related posts on social media were found to blend evidential and subjective experiential knowledge, which might result in inaccuracy [12]. A few studies examined the information quality of Reddit specifically. In [13] doctors were asked to evaluate a sample of posts related to diseases such as diabetes and AIDS on three websites, including Reddit, and found a small proportion (4 of 79) was considered as factually incorrect. Positive correlation between quality and popularity of Reddit posts has also been reported [14]. Despite the satisfactory findings, these prior studies examined the text posts only, and left out the link posts. Different from text posts, where the content is entirely in text, the link posts introduce external content by providing links and author comments. Since link posts are an important node in the path of information sharing [15], more research is needed to understand the misinformation introduced during link posting.

In this study we conducted a content analysis to examine the information quality of the Reddit link posts on health news. Since the subreddit *r/Health* does not allow text posts, and thus provides an ideal data source for studying misinformation in link posts. Glenski et al. [7] found that the majority of users in Reddit are headline browsers. They only view the summary headlines (the title in Reddit posts) and ignore the content or the comments. 73% of posts were rated without viewing the content at first. This means the quality of the summary headline written by authors in Reddit plays a key role in disseminating accurate health information on Reddit. The summary headlines are often paraphrased from the original news headlines. An examination of the types and quality of paraphrases can foster deeper understanding of the types and frequencies of inaccuracy, and thus shed light on potential strategies for curbing the misinformation dissemination in Reddit or other popular social media. Hence we focus on analyzing the quality of the paraphrases in the summary headlines. We aim to answer the following research questions:

RQ1. What are the common ways of paraphrasing when users link post health news?
RQ2. How often did inaccuracy occur in the paraphrases and what are the types of inaccuracy?

[1] https://www.redditinc.com/.

[2] https://i.imgur.com/y1Lix2T.png.

[3] https://nnlm.gov/initiatives/topics/health-literacy.

This paper is organized as follows. Section 2 offers a review of related work. In Sect. 3, the data preparation, sampling, and annotation are described, after which, in Sect. 4, the research results are discussed. In Sect. 5, the conclusions are drawn.

2 Related Work

2.1 Information Quality on Reddit Community

To date, prior research on Reddit posts has examined some aspects of the content and the community, such as the popularity of posts [16] and the structure and dynamics of the discussion forums [17]. In comparison, the information quality of the Reddit posts was less studied. Overall, a few studies have found satisfactory result regarding the information quality of Reddit posts in several topic areas. For example, Straub-Cook [18] examined the posts about public affairs in *r/Seattle*, and found that users were good at navigating and filtering the vast array of information sources. Aniche et al. [19] found that content reliability was not perceived as an issue by users in *r/programming*. A possible explanation is that since the main topics in this subreddit are about technical discussion and code sharing, the reliability of these topics may not be too difficult to assess for users with programming experience. Cole et al. [13] asked doctors to evaluate a sample of answers to health-related questions in three websites including Reddit, and found that the health information in the answers was generally accurate; only a small amount of information was assessed as poor quality. In addition, some authors explored the relationship between observed popularity and estimated quality (number of votes a post after minimizing the impact of social influence bias and inequality in visibility) in Reddit. The result shows that popularity is a relatively strong signal of quality [14]. Despite the satisfactory findings, these prior studies examined the text posts only, and left out the link posts, which is the focus of this study.

2.2 Paraphrase

Writing summary headlines after reading the articles can be regarded as a kind of paraphrase. In the area of computational linguistics, paraphrase has been well studied due to its application in information extraction [20], machine translation [21], plagiarism detection [22] and question answering [23]. Some studies have developed taxonomies for paraphrase. For example, Culicover [24] first proposes four types of paraphrase in 1968: transformational, attenuated, lexical, derivational, and real-world. Recently, Bhagat and Hovy [25] categorized the types of paraphrase into 26 categories focusing on the lexical level changes. Vila et al. [26] develop a two-tier taxonomy setting out 26 categories grouped into five classes: lexicon based changes, morphology based changes, syntax based changes, semantics based changes, and discourse based changes. Fujita [27] presents a lexical and structural paraphrase taxonomy containing six classes, namely paraphrases of single content words, function-expressional paraphrases, paraphrases of compound expressions, clause-structural paraphrases, multi-clausal paraphrases, as well as paraphrases of idiosyncratic expressions. These taxonomies were specifically designed for linguistic studies, and many defined categories

do not apply to our study of the Reddit posts. Drawing on these prior studies, we developed a simplified taxonomy that is tailored to the purpose of identifying inaccurate information. This taxonomy will be described in the next section.

2.3 Types of Inaccuracy in Health Research News

To date, researchers and media watchdogs (such as Health News Review) have been conducting manual content analysis to estimate and monitor the quality of health research news [3, 28–30]. These efforts have resulted in rich knowledge on the types of inaccuracy, especially exaggeration. Several evaluation criteria have been manually developed, such as [3, 30, 31]. Sumners et al. [3] focused on three types of exaggerations: health advice not mentioned in journal articles, causal claims from correlational findings, and human inferences from research on non-humans. Woloshin and Schwartz [32] checked the mentions of study limitations and exaggerated data presentation. Several media watchdogs, including Media Doctor Australia [28], Media Doctor Canada [29], and Health News Reviews in the United States [30], have been using a detailed 10-criteria list for media monitoring [33]. The 10 criteria used by Health News Reviews are: cost, benefit, harm, evidence, disease-mongering, funding, existing approaches, availability, novelty, and sensational language. In this study we will compare the content of the Reddit link posts and the original health research news to examine whether inaccuracy occurred during this paraphrasing process and what are the common types of inaccuracy introduced by paraphrasing.

3 Method

Our research method included multiple steps. First, we chose the Reddit health community *r/Health* as the study case and downloaded all data in the year of 2018. We then cleaned the invalid data and formed the final dataset. A sample set of posts was then selected and annotated based on our paraphrase taxonomy. Finally, we investigated the relationship between the inaccuracy and paraphrase types.

3.1 Data Preparation

The posts and related metadata in 2018 were downloaded using Pushshift.io (A website that crawls social network data in real time and open data for researchers)[4]. In total, we collected 108,235 posts. Among them were a number of advertisements and other invalid posts with no more than one comments and scores (equal to upvote minus downvote). Therefore, we removed the posts with one or zero comments and scores, and those not written in English. Finally, we obtained 4,335 valid posts.

[4] https://pushshift.io/.

To focus on the posts that may contain inaccurate paraphrases, we removed the headlines that were copied verbatim from the original news source. First, we used the python package Newspaper[5] to collect the headlines from the original news sources (A). Second, we examined whether each text posted by authors in Reddit (B) is the same as the original headline (string(A) == string(B)). In this process, 1,389 posts were found to be the same, 2,611 different, and 335 not verifiable because the original news page cannot be accessed or parsed. Therefore, the 2,611 paraphrased posts were used for developing the sample data set through random sampling. Among the paraphrased posts, three authors contributed nearly half (1,079) of the posts.

3.2 Sampling and Annotation

A sample of 250 paraphrased posts were randomly selected and annotated by one annotator. Since the posts follow the zipf law that most posts were contributed by a few authors, we selected 50 posts from each of the top 3 authors and 100 posts from all other authors. As we mentioned in Sect. 2.2, current taxonomies were specifically designed for linguistic studies, and many defined categories do not apply to our study of the Reddit posts. Hence we induced our own taxonomy based on the most obvious changes the authors made. For example, if a paraphrased post only deleted one word from original news, that post will be annotated as "Delete". The inductive coding resulted in five types of paraphrase: Copy & Paste, Combine, Add, Delete, and Replace. Table 1 lists the category and description. Table 2 shows the examples of both original sentences and paraphrased sentences in each category. Some sentences used more than one type of paraphrase.

Table 1. A taxonomy of paraphrase types and descriptions.

Paraphrase type	Description
Copy & Paste	The author copied and pasted sentences from the original news or press release
Combine	The author combined multiple original sentences together
Add	The author added their own words or sentences to the original sentence
Delete	The author deleted some of the words or phrases in the original sentence
Replace	The author replaced the words or phrases in the original sentence with their own words or phrases (also including rephrasing the whole sentences)

[5] https://github.com/codelucas/newspaper.

Table 2. Examples of sentences and paraphrase types.

Paraphrase type	Original sentence	Posted sentence
Combine	"City hosp **uses alcohol to cure** heart disease The **doctors** made use of a process known as alcohol septal ablation, in which pure alcohol is used to burn the extra mass The man was suffering from **hypertrophic cardiomyopathy**, a common disorder in which heart muscles grow thick, sometimes causing sudden death"	"Doctors use alcohol to cure hypertrophic cardiomyopathy"
Add	"Scientists Discovered What Causes Dementia"	"Scientists **Have** Discovered What Causes Dementia"
Delete	"'Raw water' is now a health trend, **because of course it is"**	"'Raw water' is now a health trend"
Replace	"A large body of evidence stretching from bench to bedside suggests that environmental stressors associated with hospitalization are toxic. Markers of allostatic overload, including elevated levels of cortisol, catecholamines, and inflammatory markers, **have been associated with** adverse outcomes after hospital discharge"	"Researchers develop the theory that the toxicity of hospitalization - lack of sleep, nutrition, activity; abundance of stress, disruption, noise, confusion - **can** have physiologic adverse effects that last long after discharge"

4 Results and Discussion

Before answering RQ1 and RQ2, we report the descriptive statistics of the health subreddit. Compared with the famous community *r/science* (21,806,873 users), the size of *r/Health* (616,042 users) is not large. To make the current participation pattern of *r/Health* clear, we collected the data of the most recent year 2018 and observed the data in two dimensions: post number and author number. Figure 1 shows the plotted numbers of both post and author. The two numbers seem to follow the same pattern, and both declined near the end of the year.

RQ1. What are the common ways of paraphrasing when users link post health news?

Since one post may use multiple types of paraphrase, we annotated each occurrence of paraphrase, and then calculated the number and percentage of posts used in each type. Table 3 shows that "copy and paste" was the most common type of paraphrase; it was used in 121 posts (48.4%). "Replace" was also frequently used (32.4%). "Add" and "Delete" were less common (13.2% and 22.8% respectively). Table 4 shows examples of common types of paraphrase.

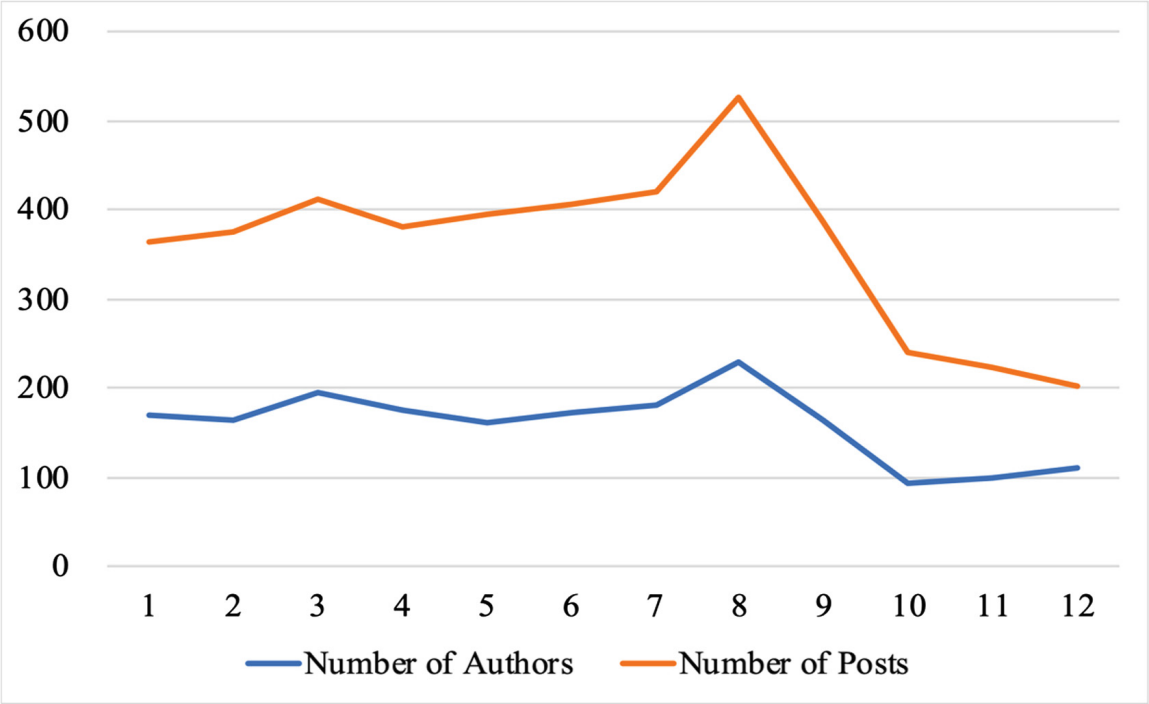

Fig. 1. Trend of the Reddit health community by month

Table 3. Paraphrase type distribution

Paraphrase type	Count	Percentage
Copy & Paste	121	48.4%
Combine	26	10.4%
Add	33	13.2%
Delete	57	22.8%
Replace	81	32.4%

Table 4. Examples of common types of paraphrase

Paraphrase type	Example
Add	**Time** (in the future), **Location** (in America), **Sources** (finds new research, Science, PubMed, in a new study, Interesting study), **Triggering discussion** (is there a downside), **Function words** (have), **Content words** (alone, deep, disease, neurocognitive scores)
Delete	**Time** (Tuesday), **Sources** (study says, Gallup found), **Function words** (but, and, a, its, their)
Replace	**Simplicity** (and = &, two = 2, percent = %, United States = US, administration officials = admin, The tattoo and the hospital's decision = It), **Personal habits** (said = told, predicted = suggests, favored = vote for, recently = just), **Certainty** (may = can, may = activated, would = could)

RQ2. How often did inaccuracy occur in the paraphrases and what are the types of inaccuracy?

Drawing on the categorization of inaccuracy in prior studies [3], we categorized the 12 posts with inaccurate information into four types: *exaggerated causal claims* (7 posts), *exaggerated inference to human or larger population (*1 post*), unconfirmed new claim (*1 post*)*, and *other factual errors* (3 posts).

Exaggerated causal claim is a major type of misinformation in health research press releases and news stories [3]. In our sample data, we found seven posts containing exaggerated causal claims. In these paraphrases, the authors often removed the modal verbs (such as "may" and "might") that were used to mitigate the certainty in the original news, resulting in increased certainty and exaggerated claims. For example, in Ex. 1, the original news on a potential link between the *Tau* protein and Alzheimer's diseases was exaggerated as if it is a direct causal finding by removing the modal verbs and replacing the original phrase of "strong link" with a direct causal verb "activate". Since this claim and other relevant claims on the cause of Alzheimer's disease [34] are extremely important for finding treatments for Alzheimer's disease, which is affecting 44 millions of people in the world, the exaggerated claims may lead to false hope for researchers and patients.

Ex. 1 Original News: New evidence suggests a mechanism by which progressive accumulation of Tau protein in brain cells **may** lead to Alzheimer's disease. Scientists studied more than 600 human brains and fruit fly models of Alzheimer's disease and found the first evidence of a **strong link** between Tau protein within neurons and the activity of particular DNA sequences called transposable elements, which **might** trigger neurodegeneration.

Paraphrased Post: Tau **Activates** Transposable Elements in Alzheimer's Disease.
In the other cases of exaggerated causal claims, "may" was replaced by "can", "have been associated with" replaced by "can", "associated with" replaced by "as a result", "would" replaced by "could", "would prevent" replaced by "prevents", and "could reduce the number" replaced by "without", indicating exaggeration from correlational or conditional causal findings to direct causal claims.

Exaggerated inference from animal studies to humans, or from small samples to larger population has been found in previous studies [1]. In our data set we found only one post that removed important research details, resulting in exaggerated inference from a small group of patients to all patients. In Ex. 2, the paraphrased post removed the information that the new treatment will be tested on three patients only, and large-scale clinical trial is expected in the future. Since study design [35] is considered important information for understanding the strength of the research findings, removing the detail on the small sample size could result in a false belief that the treatment is available to all patients.

Ex. 2. Original News: Scientists in Japan now have permission to treat people who have heart disease with cells produced by a revolutionary reprogramming technique. On 16 May, Japan's health ministry gave doctors the green light to take wafer-thin sheets of tissue derived from iPS cells and graft them onto diseased human hearts. In their technique, Sawa and his colleagues use iPS cells to create a sheet of 100 million

heart-muscle cells. Once Sawa's team has treated its **three patients**, it will apply to conduct a **clinical trial** involving a further seven to ten people.

Paraphrased Post: Scientists in Japan now have permission to treat people who have heart disease with tissue derived from human induced pluripotent stem cells. Sheets of up to 100 million heart muscle cells grown in a lab will be surgically applied to diseased hearts.

Another post (Ex. 3) combined two findings, one on children's yoghurts and the other on organic yoghurts into a new, unconfirmed claim on "children's organic yoghurts".

Ex. 3. Original News: In our survey of yogurts sold in the UK, we found that less than 10% were low sugar – almost none of which were **children's yogurts**. We also found that **organic** products, often viewed as healthier options, contained some of the highest levels of sugar.

Paraphrased Post: Organic children's yogurts found to have some of the highest sugar contents in the product line.

Other factual errors were also found. In Ex. 4, the author of the following paraphrase misunderstood the meaning of "genomes" and replaced the phrase "1 million or more volunteers" with "1 million genomes", indicating the author's lack of biomedical knowledge. In Ex. 5, one sentence not found in the original news was added, inserting unconfirmed information. In another post, "$60" was replaced by "$37", probably a typo.

Ex. 4. Original News: NIH launches All of Us research program this week. The National Institutes of Health (NIH) announced this week plans to open national enrollment for the All of Us Research Program on May 6. The goal of the program is to **enroll 1 million or more volunteers**. Through the program, the volunteers will agree to share information about themselves over many years. "All of Us is an ambitious project that has the potential to revolutionize how we study disease and medicine," Health and Human Services Secretary Alex Azar said.

Paraphrased Post: The US is launching a massive effort to **sequence 1 million genomes** and link them to personal health in order to better study disease and medicine.

Ex. 5. Original News: A new, eye-wateringly high estimate of the cost of obesity in the US. A report released this week puts a surprisingly high figure on the societal cost of obesity in the US: $1.72 trillion annually, or 9.3% of GDP. By far the biggest chunk of that $1.72 trillion is the $1.24 trillion chunk attributed to the "indirect" costs of obesity: the "work absences, lost wages, and reduced economic productivity for the individuals suffering from the conditions and their family caregivers," the report explains. That is, the bulk comes from costs other than healthcare spending.

Paraphrased Post: A new study estimates the obesity estimate costs $1.7 trillion a year in the US alone. Or more than $5,000 per person per year. Increased risk of **arthritis, back/knee pain, early disability, early retirement, diabetes, heart disease, cancer** all drive this cost.

The RQ2 result shows that although most link posts were faithful to the original news content, a small number of posts contained distorted information regarding the major research finding, study design, and other factual information. The most common type of misinformation is the exaggerated causal claims, which account for more than half of the problematic posts. The exaggerated causal claims, exaggerated inference to larger population, and new claim invented with no evidence may result in misunderstanding and false hope for researchers and patients.

5 Conclusions

In this study, we sampled 250 link posts from the Reddit health community *r/Health* in 2018, and examined the inaccurate information introduced when authors paraphrased the original health news stories. The result shows that most posts simply copied the original news content verbatim, while some paraphrased by adding, deleting, replacing, and combining content. We found a total of 12 paraphrased posts that contained inaccurate information. The most common type of inaccuracy is exaggeration resulted from changing the original speculative claims to direct causal statements by removing the modal verbs such as "may" and "might". Although the link posts were generally consistent with the original news stories, the small number of exaggerated claims may lead to false hope for researchers and patients. This study provides a first-step information quality scan of the link posts on Reddit. Due to the limited amount of data included in this study, a larger-scale study is needed to be able to generalize the finding to the broader community of volunteers for science communication on social media.

References

1. Tankard, J.W., Ryan, M.: News source perceptions of accuracy of science coverage. Journal. Q. **51**, 219–225 (1974). https://doi.org/10.1177/107769907405100204
2. Pellechia, M.G.: Trends in science coverage: a content analysis of three US newspapers. Public Underst. Sci. **6**, 49–68 (1997). https://doi.org/10.1088/0963-6625/6/1/004
3. Sumner, P., et al.: The association between exaggeration in health related science news and academic press releases: retrospective observational study. BMJ **349**, g7015 (2014). https://doi.org/10.1136/bmj.g7015
4. Chang, C.: Inaccuracy in health research news: a typology and predictions of scientists' perceptions of the accuracy of research news. J. Health Commun. **20**, 177–186 (2015). https://doi.org/10.1080/10810730.2014.917746
5. Fahnestock, J.: Accommodating science: the rhetorical life of scientific facts. Writ. Commun. **15**, 330–350 (1998). https://doi.org/10.1177/0741088398015003006
6. Buhse, S., Rahn, A.C., Bock, M., Mühlhauser, I.: Causal interpretation of correlational studies – analysis of medical news on the website of the official journal for German physicians. PLoS ONE **13**, e0196833 (2018). https://doi.org/10.1371/journal.pone.0196833
7. Glenski, M., Pennycuff, C., Weninger, T.: Consumers and curators: browsing and voting patterns on reddit. IEEE Trans. Comput. Soc. Syst. **4**, 196–206 (2017). https://doi.org/10.1109/TCSS.2017.2742242

8. Brossard, D., Scheufele, D.A.: Science, new media, and the public. Science **339**, 40–41 (2013). https://doi.org/10.1126/science.1232329
9. Ovadia, S.: More than just cat pictures: Reddit as a curated news source. Behav. Soc. Sci. Libr. **34**, 37–40 (2015). https://doi.org/10.1080/01639269.2015.996491
10. Zhao, Y., Zhang, J.: Consumer health information seeking in social media: a literature review. Health Inf. Libr. J. **34**, 268–283 (2017). https://doi.org/10.1111/hir.12192
11. de Belt, T.H.V., Engelen, L.J., Berben, S.A., Teerenstra, S., Samsom, M., Schoonhoven, L.: Internet and social media for health-related information and communication in health care: preferences of the Dutch general population. J. Med. Internet Res. **15**, e220 (2013). https://doi.org/10.2196/jmir.2607
12. Sharma, R., Wigginton, B., Meurk, C., Ford, P., Gartner, C.: Motivations and limitations associated with vaping among people with mental illness: a qualitative analysis of Reddit discussions. IJERPH **14**, 7 (2016). https://doi.org/10.3390/ijerph14010007
13. Cole, J., Watkins, C., Kleine, D.: Health advice from internet discussion forums: how bad is dangerous? J. Med. Internet Res. **18**, e4 (2016). https://doi.org/10.2196/jmir.5051
14. Stoddard, G.: Popularity dynamics and intrinsic quality in Reddit and hacker news. In: ICWSM (2015)
15. Record, R.A., Silberman, W.R., Santiago, J.E., Ham, T.: I sought it, i Reddit: examining health information engagement behaviors among Reddit users. J. Health Commun. **23**, 470–476 (2018)
16. Horne, B.D., Adali, S., Sikdar, S.: Identifying the social signals that drive online discussions: a case study of Reddit communities. In: 2017 26th International Conference on Computer Communication and Networks (ICCCN), pp. 1–9 (2017)
17. Medvedev, A.N., Delvenne, J.-C., Lambiotte, R.: Modelling structure and predicting dynamics of discussion threads in online boards. J. Complex Netw. **7**, 67–82 (2019). https://doi.org/10.1093/comnet/cny010
18. Straub-Cook, P.: Source, please? Digit. Journal. **6**, 1314–1332 (2018). https://doi.org/10.1080/21670811.2017.1412801
19. Aniche, M., et al.: How modern news aggregators help development communities shape and share knowledge. In: 2018 IEEE/ACM 40th International Conference on Software Engineering (ICSE), pp. 499–510 (2018)
20. Shinyama, Y., Sekine, S.: Paraphrase acquisition for information extraction. In: Proceedings of the Second International Workshop on Paraphrasing, vol. 16, pp. 65–71. Association for Computational Linguistics, Stroudsburg (2003)
21. Callison-Burch, C., Koehn, P., Osborne, M.: Improved statistical machine translation using paraphrases. In: Proceedings of the Main Conference on Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics, pp. 17–24. Association for Computational Linguistics, Stroudsburg (2006)
22. Barrón-Cedeño, A., Vila, M., Martí, M., Rosso, P.: Plagiarism meets paraphrasing: insights for the next generation in automatic plagiarism detection. Comput. Linguist. **39**, 917–947 (2013). https://doi.org/10.1162/COLI_a_00153
23. Fader, A., Zettlemoyer, L., Etzioni, O.: Paraphrase-driven learning for open question answering. In: Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics, Long Papers, vol. 1, pp. 1608–1618. Association for Computational Linguistics, Sofia (2013)
24. Culicover, P.W.: Paraphrase generation and information retrieval from stored text. Mech. Transl. Comput. Linguist. **11**, 78–88 (1968)
25. Bhagat, R., Hovy, E.: What is a paraphrase? Comput. Linguist. **39**, 463–472 (2013). https://doi.org/10.1162/COLI_a_00166

26. Vila, M., Martí, M.A., Rodríguez, H.: Paraphrase concept and typology. A linguistically based and computationally oriented approach. Procesamiento del Lenguaje Nat. **46**, 83–90 (2010)
27. Fujita, A.: Automatic generation of syntactically well-formed and semantically appropriate paraphrases (2005)
28. Smith, D.E., Wilson, A.J., Henry, D.A.: Monitoring the quality of medical news reporting: early experience with media doctor. Med. J. Aust. **183**, 190–193 (2005). https://doi.org/10.5694/j.1326-5377.2005.tb06992.x
29. How well do Canadian media outlets convey medical treatment information? https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3090174/
30. Schwitzer, G.: How do US journalists cover treatments, tests, products, and procedures? An evaluation of 500 stories. PLOS Med. **5**, e95 (2008). https://doi.org/10.1371/journal.pmed.0050095
31. Chang, C.: Inaccuracy in health research news: a typology and predictions of scientists' perceptions of the accuracy of research news. J. Health Commun. **20**, 177–186 (2015)
32. Woloshin, S., Schwartz, L.M.: Press releases: translating research into news. JAMA **287**, 2856–2858 (2002)
33. Moynihan, R., et al.: Coverage by the news media of the benefits and risks of medications. N. Engl. J. Med. **342**, 1645–1650 (2000)
34. Greenberg, S.A.: How citation distortions create unfounded authority: analysis of a citation network. BMJ **339**, b2680 (2009)
35. Cyranoski, D.: 'Reprogrammed' stem cells approved to mend human hearts for the first time. Nature **557**, 619 (2018)