Exploring rare disease realities: a systematic literature review on harnessing social media patient-generated data

Emma Le Priol, Anaïs Gedik, Adel Mebarki, Stéphane Schück, Nathalie Texier, Anita Burgun

Exploring rare disease realities: a systematic literature review on harnessing social media patient-generated data

Emma Le Priol, Anaïs Gedik, Adel Mebarki, Stéphane Schück, Nathalie Texier, Anita Burgun

Submitted to: Journal of Medical Internet Research
on: September 08, 2023

Table of Contents

Exploring rare disease realities: a systematic literature review on harnessing social media patient-generated data

Emma Le Priol[1, 2]; Anaïs Gedik[2]; Adel Mebarki[2]; Stéphane Schück[2]; Nathalie Texier[2]; Anita Burgun[1, 3, 4]

[1]HeKA team INSERM UMR 1138 Université Paris-Cité Paris FR
[2]Kap Code Paris FR
[3]Clinical Bioinformatics group Institut Imagine Paris FR
[4]Department of Medical informatics Necker Hospital AP-HP Paris FR

Corresponding Author:
Emma Le Priol
Kap Code
146 rue Montmartre
Paris
FR

Abstract

Background: Rare diseases affect roughly 400 million people world-wide, and 30 million in Europe. Intangible costs and personal aspects for patients and their families are rarely accounted for, as most studies focus on easy-to-measure metrics. On the other hand, social media play an important role for these people who can easily feel isolated and seek both support and advice online.

Objective: We aim to examine in the peer-reviewed academic literature how social media has been used to generate new knowledge, and the types of research questions that have been answered through social data mining. We explore what types of methods, data sources, and data types are used.

Methods: We reviewed studies based on user-generated data, focusing on rare diseases, and published prior to May 2023. For included publications, a list of pertinent variables was established to cover data sources, data processing, and objectives. These variables were later on analysed quantitatively and qualitatively.

Results: Eighty-seven studies were included. The vast majority of publications (94.3%) focused on one rare disease or on a family of rare diseases. Overall, only less than a hundred rare diseases were studied in the included publications. Moreover, 93.1% of the studies analysed contents in English. Surprisingly, automated methods were used in only seven studies, all published after 2020. These publications' mean number of posts studied is 33,201 (compared to 1,405 for publications analysing the posts manually). Among these publications, three had a temporal range of five years or more, accounting for half of the publications with a temporal range of five years or more (the majority of publications had a temporal range of less than two years). Among the seven publications using AI methods, the two main AI-assisted tasks were sentiment analysis and topic identification.

Conclusions: This work allowed us to grasp what the reality of using user-generated social media data for rare disease research was in 2023. The opportunities of current AI research on NLP are still underexploited in this very specific field, resulting in an under exploitation of online data. Contrasting with the high expectancies of the rare disease research community, this review shows that social media based studies in this field are still at an early stage, with only a tiny portion of rare diseases studied, with only a few languages studied also, and mainly with only very few studies exploiting current NLP progress to extract knowledge from social media data.

(JMIR Preprints 08/09/2023:52568)
DOI: https://doi.org/10.2196/preprints.52568

Preprint Settings

1) Would you like to publish your submitted manuscript as preprint?

✔ **Please make my preprint PDF available to anyone at any time (recommended).**
Please make my preprint PDF available only to logged-in users; I understand that my title and abstract will remain visible to all users.

Only make the preprint title and abstract visible.
No, I do not wish to publish my submitted manuscript as a preprint.

2) If accepted for publication in a JMIR journal, would you like the PDF to be visible to the public?

✓ **Yes, please make my accepted manuscript PDF available to anyone at any time (Recommended).**
Yes, but please make my accepted manuscript PDF available only to logged-in users; I understand that the title and abstract will remain v
Yes, but only make the title and abstract visible (see Important note, above). I understand that if I later pay to participate in <a href="http

Original Manuscript

Review

Exploring rare disease realities: a systematic literature review on harnessing social media patient-generated data

Abstract

Background: Rare diseases affect roughly 400 million people world-wide, and 30 million in Europe. Intangible costs and personal aspects for patients and their families are rarely accounted for, as most studies focus on easy-to-measure metrics. On the other hand, social media play an important role for these people who can easily feel isolated and seek both support and advice online.
Objective: We aim to examine in the peer-reviewed academic literature how social media has been used to generate new knowledge, and the types of research questions that have been answered through social data mining. We explore what types of methods, data sources, and data types are used.
Methods: We reviewed studies based on user-generated data, focusing on rare diseases, and published prior to May 2023. For included publications, a list of pertinent variables was established to cover data sources, data processing, and objectives. These variables were later on analysed quantitatively and qualitatively.
Results: Eighty-seven studies were included. The vast majority of publications (94.3%) focused on one rare disease or on a family of rare diseases. Overall, only less than a hundred rare diseases were studied in the included publications. Moreover, 93.1% of the studies analysed contents in English. Surprisingly, automated methods were used in only seven studies, all published after 2020. These publications' mean number of posts studied is 33,201 (compared to 1,405 for publications analysing the posts manually). Among these publications, three had a temporal range of five years or more, accounting for half of the publications with a temporal range of five years or more (the majority of publications had a temporal range of less than two years). Among the seven publications using AI methods, the two main AI-assisted tasks were sentiment analysis and topic identification.
Conclusions: This work allowed us to grasp what the reality of using user-generated social media data for rare disease research was in 2023. The opportunities of current AI research on NLP are still underexploited in this very specific field, resulting in an under exploitation of online data. Contrasting with the high expectancies of the rare disease research community, this review shows that social media based studies in this field are still at an early stage, with only a tiny portion of rare diseases studied, with only a few languages studied also, and mainly with only very few studies exploiting current NLP progress to extract knowledge from social media data.
Keywords: rare diseases; genetic rare diseases; congenital diseases; social media; natural language processing; artificial intelligence; systematic review; real-world data

Introduction

Background

Rare diseases are defined in Europe by a prevalence of less than 5 for 10,000 inhabitants. Although each of these diseases affects a small number of people, between 5,000 and 8,000 rare diseases have been identified world-wide which globally affect roughly 400 million people, with approximately 30 million in Europe [1]. However, due to the general lack of knowledge and the little probability for a doctor to have already crossed paths with the rare disease corresponding to the symptoms, patients wait on average one year and a half before being diagnosed. And for about one fourth, it takes more than five years before getting the right diagnosis and being able to be treated [2]. Most of the time, the patient's journey ends in a situation that negatively impacts their health and psychosocial wellbeing, as well as that of their families [3,4].

Eighty percent of rare diseases have a genetic origin, and seventy percent start in childhood [5–7]. These diseases may be associated with reduced life expectancy, and physical symptoms and disabilities that can compromise activities, autonomy, and well-being. They can also have psychological as well as social impacts and costs for individuals, families, healthcare systems and society as a whole. There have been some attempts to assess the burden of rare diseases. The EveryLife Foundation investigated the economic burden of 379 rare diseases in the US and showed that excess costs were associated with hospital inpatient care, and labour market productivity losses due to absenteeism and early retirement [8].

Most studies focus on easy-to-measure economic costs (e.g., in/outpatient care, drugs, absenteeism, etc) or specific social determinants like work participation [9], while more intangible costs and personal aspects are usually not addressed [10]. With the objective of collecting patients' opinions and transforming patients' and families' experiences into facts and figures that can be shared with a wider public and policymakers, the Juggling Care Survey was conducted in 2017 by the European rare disease patient association EURORDIS [11]. This online survey highlighted patients' and families' needs, impacts on daily life for patients and their families, difficulties in coordination of care, in access to services, and work-life balance, as well as impacts on well-being and mental health. They showed that the disease has a severe or very severe impact on everyday life for more than half of the patients and "severe effects on social and family life, thus triggering isolation and feelings of being neglected for some members of the family".

Social media play a crucial role for rare disease patients and families. They use social media to tell their stories, share knowledge about their diseases and symptoms, ask for advice and look for community help as well as develop advocacy plans to get funding and greater awareness. By definition, rare diseases can easily cause patients and their families to feel isolated. This is even more important for ultra-rare diseases – i.e. conditions with prevalence < 1/50,000. Social media break down the geographical barriers and more and more rare disease patients are using digital platforms to share their experience and look for support. For example, the frustration and lack of answers lead patients or their families to seek support and answers online, and more particularly on social media [12]. With 63% of the world's population using the internet in 2021 [13], and 93.4% of internet users being on social media [14], social media are growing into really rich sources of information not only for individuals – patients and families – but also for public health policymakers.

Prior work

Prior work by Miller et al. [15] provided a broad overview of the uses of social media in rare disease research. They analysed all possible uses of social media in research, for example patient recruitment for clinical trials, dissemination of surveys, etc. Their review focused on articles indexed before November 2020 whereas digital health and patient-driven strategies have mostly been implemented in the last couple of years. They showed that despite the potential benefits of studying social media for rare disease research, there are still some methodological limits. However, they did not specifically study contents generated by the patients, their families or professionals nor possible methods to automate the analysis using Artificial Intelligence (AI) technologies.

Objective

The goal of this study is to systematically review the peer-reviewed academic literature on the use of social media data generated by rare disease patients and families. More precisely, our study will be focused on genetic and congenital rare diseases. In this review, we examine how social media has been used to generate new knowledge, and the types of research questions that have been answered

through social data mining. Moreover, we explore what types of methods, data sources, and data types are used, with a focus on major results, uncovered needs and opportunities in rare-disease-oriented AI research using social media.

Methods

Search strategy

Starting from the search strategy used by Miller et al. [15], we extended the coverage up until May 2023 and limited ourselves to publications based on user-generated data.
The first step was to integrate in our corpus publications classified as “content analysis” in Miller et al.’s review [15].

Then, based on their search strings (see appendix A [15]), five databases (PubMed, Embase, CINAHL, PsycINFO and Web of Science) were queried for the period going from November 2020 to May 2023.

Selection of relevant publications

Publications were included if they met the following criteria: (i) focused on rare diseases, (ii) rare communicable diseases (such as Monkeypox disease) were excluded, (iii) used social media data as a primary source of data for their research (studies using social media to collect other types of data were excluded), (iv) were published in English in a peer-reviewed journal between January 2004 and May 2023. The review of all publications was made by ELP and in case of uncertainty, our protocol was to get another review by AB, then a consensus was reached.

Because publications might be indexed by more than one database, duplicates were removed.

Identification of relevant variables

A list of variables was established to annotate the selected articles. They cover data sources, data processing (including potential AI-methods), and objectives. They can be grouped into five categories:

- Metadata associated with the publication: title, DOI, authors, journal, year of publication
- Objectives of the publication: study goal as stated in the publication, as well as the category under which it falls (c.f. Subsection study goals of the results), disease(s) studied and whether it is a single rare disease, a group of diseases or rare diseases as a whole
- Data source: social media from which data is extracted, language of the posts, temporal range (i.e. on what period of time were the messages studied posted), number of posts included, number of users included
- Methods and results: methods and results as stated in the publication
- AI model: when relevant, purpose of the model, type of model used, performance score used, and its value

Data extraction and analysis

The results were exported into Zotero and all the articles were reviewed and annotated based on the variables listed above. The results were stored into an Excel sheet provided as supplementary material.

The variables created were analysed both quantitatively and qualitatively.

Results

The first result is the small number of publications meeting our inclusion criteria (13.9%) compared to the number of publications retrieved from the query on the five databases.

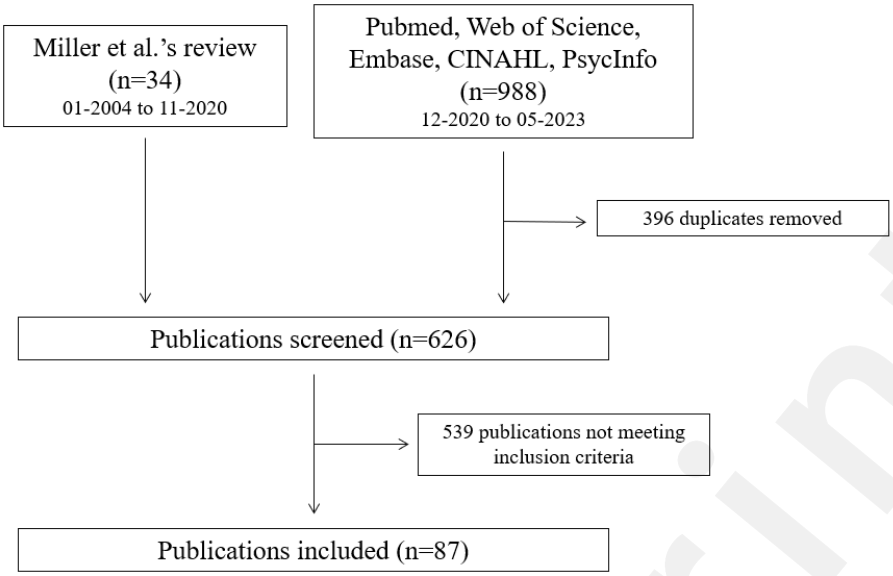


Figure 1. Flow diagram.

The publications not included give an overview of the other uses of social media in rare disease research, e.g. participant recruitment in studies and clinical trials through social media (n=87), use of social media to survey a given population (n=149), evaluation of fundraising or awareness campaigns (n=17), use of social media (in particular WeChat) to remotely follow-up on patients after a hospitalisation or to give pre-hospitalisation information (n=16).

Another finding is the increasing number of publications in the early 2020s. Figure 2 shows that it is fairly recent that researchers have begun to show interest in rare disease user-generated data. While the number of publications per year meeting our criteria is less than one before 2014 and less than five between 2014 and 2020, we extracted 21 articles in 2021, 29 in 2022 and 9 more were indexed between January and May 2023. Globally, more than two thirds (n=59; 67.8%) of the publications were indexed between January 2021 and May 2023.

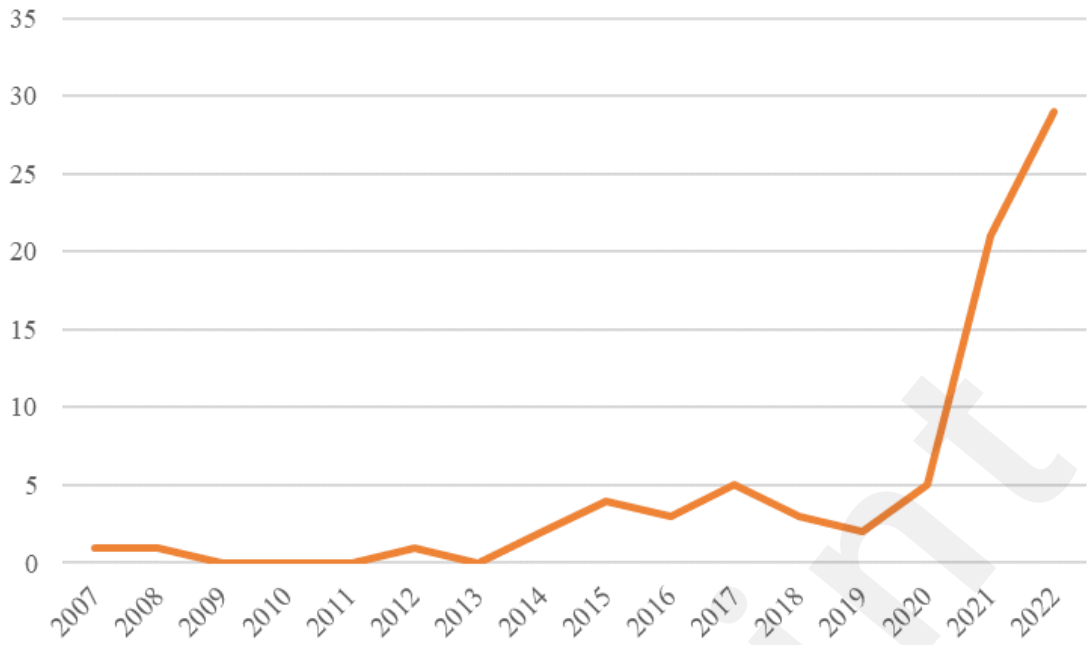


Figure 2. Number of included publications by year of publication.

Rare diseases

The vast majority of studies (94.3%) focused on one rare disease (n=60; 69.0%) or on a family of rare diseases (n=22; 25.3%). Studies that encompass multiple rare diseases were relatively rare (n=5; 5.7%).

Table 1. Number and proportion of publications by focus.

	Number of studies	Percentage of studies
Single rare disease	60	69.0
Group of rare diseases	22	25.3
Comparison of rare diseases	1	1.1
Rare diseases in general	4	4.6

Besides Sickle Cell Disease (n=6; 6.9%), Cleft Lip/Palate (n=5; 5.7%), Cystic Fibrosis (n=4; 4.6%) - which represent almost a fifth of the publications (17.2%), the other diseases were mentioned in only one (n=19; 21.8%) or two (n=5; 5.7%) publications each. Twenty-two publications studied groups of diseases, encompassing two or more rare diseases each. Overall, 44 different rare diseases and 20 groups were studied, which amounts to less than 100 different rare diseases in total. Only four publications [16–19] considered rare diseases in general and only one [20] compared the online activity about two different diseases.

Study goals

Categories and subcategories were defined to classify the publications according to their focus:

- on patients or their families: the subjects were varied and ranged from the exploration of thematics mentioned online; to the description of reported experience of the disease and the needs and concerns expressed, the treatment or other specific aspects of the disease such as bullying for example [21]; or to the identification on social media of rare disease patients or finally to the impact of Covid-19 pandemic on them.
- on online groups and/or communities

- on the scientific relevance and/or quality of online contents
- on patients' and professionals' contents
- other

Table 2. Number and proportion of publications by study goal.

	Number of studies	Percentage of studies
Patients and their families	39	44.8
Categorisation of topics	13	14.9
Experience of the disease/treatment	11	12.6
Needs and concerns	6	6.9
Identification of patients	2	2.3
Impact of Covid-19	4	4.6
Evaluating online support	3	3.4
Communities and groups	18	20.7
Scientific relevance/quality of online content	18	20.7
Comparison between patients' and professionals' contents	3	3.4
Other	9	10.3

Patients or their families (44.8%)

Almost half (44.8%) of the publications had goals related to the patients or their families. These goals encompass a wide range of topics.

Patients' and caregivers' insights

Many of the publications aimed at better understanding patients' or caregivers' experience, point of view, and quality of life. These publications usually did not focus on only one aspect and could include insights about the disease experience, the symptoms, the quality of life, the care and treatment pathways, the diagnosis, the seeking and giving of advice, the disease awareness [22–32].

Among them, one publication reported on using social media interactions to develop a comprehensive disease model. Goodspeed et al. developed a draft conceptual model of SLC6A1 Neurodevelopmental Disorder (SLC6A1-NDD), a disease that was first described in 2015 [26]. This draft conceptual model, which aimed at aiding in the design of natural history studies, was based on all cases reported in the published literature, on interviews of two key opinion leaders and on an analysis of interactions between caregivers on social media [26]. This kind of paper is a good example of the use of social media data conjointly to other sources to crack open the natural history of a rare and relatively unknown disease.

Insights about specific experiences

However, some publications tackled more specific aspects of patients' and caregivers' experiences.
For example, Stewart et al. used a research-specific online forum to understand feeding difficulties of children following esophageal atresia/tracheo-esophageal fistula repair and the impact of these feeding difficulties on parents [33]. They showed that beyond the child's health and development,

feeding difficulties had a huge impact on the parents' well-being and quality of life (anxiety, trauma, uncertainty and isolation) [33]. Similarly, Goodspeed et al. not only confirmed that the core symptoms of epilepsy and autistic traits were prominent concerns in SLC6A1-NDD but also demonstrated that other symptoms have a large impact on family life [26].

Korkmaz et al. provide another example of a publication tackling a very specific aspect of a congenital disease, Cleft Lip/Palate (CLP): bullying [21]. They used Twitter data to evaluate bullying in individuals with CLP and showed that "most tweets posted by individuals with CLP and their relatives were about their personal experiences of being bullied and how it affected their lives" but that there were also tweets of social support against bullying and of news about bullying in CLP [21].

These examples suggest that these studies may highlight differences between the aspects of disease reported in the literature and those discussed in patient conversations.

Insights specifically about treatments

While disease treatments were usually included in publications that have overall objectives of understanding patients' experiences, some publications focused solely on this specific aspect.

For example, Walker et al. had the objective to better understand barriers to Hydroxyurea (HU) for Sickle Cell Disease (SCD) patients [34]. To achieve that goal, they conducted a study of Facebook messages posted in a support group [34]. The first conclusion drawn was that 35% of the users expressed supportive statements towards HU compared to 25% who expressed non supportive statements [34]. They also found an emergent theme in the Facebook discussions with some patients arguing that HU may "mask" SCD complications rather than improving them [35].

Similarly, Kline et al. tackled barriers to genomic medicine access for those living with Ehlers-Danlos syndrome and hypermobility spectrum disorders [36]. Their results confirmed the barriers documented in previous studies [36]. They also found that the barriers were mostly social-structural and interpersonal with rates 10%–47% higher than other levels of influence [36]. Their main contribution is showing that social media data allowed searchers to grasp barriers that patients and caregivers wouldn't have considered as barriers in an interview or a survey [36].

Other papers tackled the topic of patients' perspectives on treatments, especially for heavy surgical treatments such as tooth removal for Lesch-Nyhan disease patients [37]. Another example is the study of patients' perspectives on the symptomatic treatments of Amyotrophic Lateral Sclerosis, which lack randomised controlled trials and therefore lack information about safety, efficacy and side effects [38]. Similarly, new treatments can be studied: for example, Mahoney et al. studied the impact on parents' hopes and expectations of a new treatment for Duchenne muscular dystrophy or spinal muscular atrophy type 2 [39].

Insights about genetic risk and testing

In the same way as treatments, genetic risk and testing were often encompassed in publications with broad objectives, but some publications specifically focused on this topic.

Howard et al. explored how people with a family history of Motor Neuron Disease (MND) also known as Amyotrophic Lateral Sclerosis make sense of and negotiate genetic risk, based on MND Association's online forum exchanges [40]. They showed that the forum was a space for sharing personal experiences, knowledge and information, helping forum users understanding and acting upon genetic risk [40]. They also showed that people with familial history of MND develop an important awareness of the risk and gave a particular significance to symptoms that other people

might dismiss quite easily [40]. Presymptomatic genetic testing was also an important topic on the forum, often linked with discussions about starting a family and having children [40].

In a similar way, Smedley and Coulson analysed health forums posts about genetic testing of Huntington's disease [41]. They showed that three themes stand out: deciding to be tested, preparing for the test and receiving the results [41].

Insights regarding the impact of Covid-19

Of note, four articles (4.6%) studied the consequences of Covid-19 on patients' [42–44] or parents' [45] experiences and concerns. These publications focused on the disruption caused by the mondial crisis for people suffering from rare rheumatologic, cardiac, pulmonary, or esophageal diseases.

Evaluating support on social media

Patients' and caregivers' messages were also used in scientific literature to assess the support given and received through social media. For example, Coulson et al. analysed messages from a Huntington's disease online support group using a social support framework and showed that informational and emotional support were the most offered, while esteem support and tangible assistance were more rarely offered [46].

Communities and groups (20.7%)

The studies in this category aimed at understanding the group dynamics in the context of a specific rare disease. Even if the data used was still at user-level, the focalisation was on the community. One example is Wittmeier et al.'s publication, in which they analysed the use of a social media community for Hirschsprung's Disease [47]. They showed that such a community was used for discussion, support and advocacy ; and that overall it played an important role in connecting families [47].

Scientific relevance/quality of online content (20.7%)

The publications in this category were all very similar, and aimed at evaluating the quality of online content about a rare disease or a treatment. They were all but one [48] based on YouTube videos. Overall, they concluded that videos were useful for the patients and their families but they pointed out limits. The main limit was the insufficient quality of some videos [49–57]. But other limits were found such as the lack of videos [58], the lack of educational content for non-professional viewers or the need to have a medical background to understand the videos [54,55].

Comparison between patients' and professionals' contents (3.4%)

This category regroups publications that aimed to compare patients' and healthcare professionals' opinions and uses of social media.

For example, Henrick et al. compared how professionals and patients discussed Amelogenesis Imperfecta on eight social media platforms [59]. They showed that these platforms were not only places for patients and caregivers to share their experiences, but also for professionals to share and seek information, suggesting that both patients and professionals lack proper information about Amelogenesis Imperfecta and seek it online [59].

Social media

Interestingly, almost half (n=41; 47.1%) of the publications did not focus on a single social medium but on more than one. For those which focused on only one social media, the most frequent were YouTube (n=16; 18.4%), Twitter (n=11; 12.6%) and Facebook (n=10; 11.5%). Other social media studied in the publications included Instagram, TikTok, Baidu Tieba, as well as blogs and forums

dedicated to a given rare disease. The most studied social media overall were Twitter (n=27; 31.0%), Facebook (n=27; 31.0%) and YouTube (n=23; 26.4%). TikTok, which is a relatively new social media, was rarely studied alone (n=2; 2.3%).

Table 3. Number and proportion of publications by social media[a].

	Number of studies	Percentage of studies
Twitter	27	31.0
Facebook	27	31.0
YouTube	23	26.4
Forum/blog	18	20.7
Instagram	17	19.5
Reddit	11	12.6
TikTok	9	10.3
Not precised/other	4	4.6

Type of content

Except for one publication [60] – in which YouTube videos as well as comments were used (it was included in the video category)– all studies focused on only one type of content (text, image, or video).

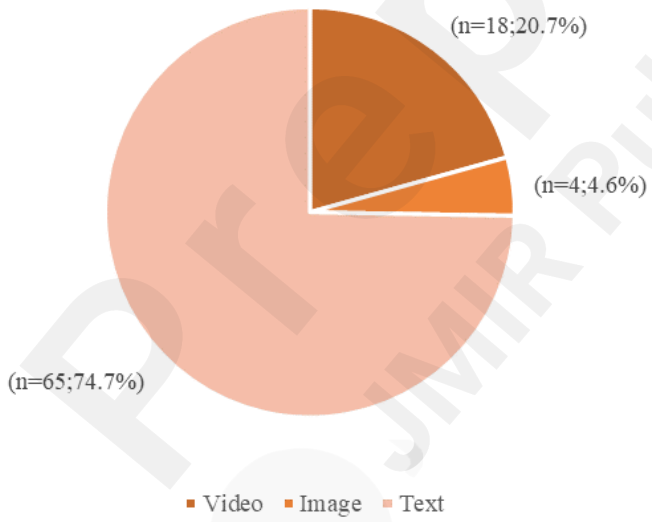


Figure 3. Type of content analysed.

Depending on the social media, the type of content studied in the publication also varied. If text was the most studied type of content (n=65; 74.7%), video, probably with the increasing use of video-based platforms such as YouTube and TikTok, seemed to be the subject of a growing proportion of publications (n=18; 20.7%). Of note, the number of studies using video has increased in the last three years: among the eighteen publications focused on video, all but one [61] were published after 2020. Similarly, the four publications focused on images [62–65] were all published after 2021.

Languages of the content

In the vast majority of the studies (n=81; 93.1%), the contents analysed were in English. Other languages are French (n=3; 3.4%), Spanish (n=2; 2.3%), and Danish, Dutsch, German and Chinese

[a] Categories are not exclusive, so the sum of % can be more than 100%.

(only once each).

Methods used to conduct the studies

Surprisingly, automated methods were used in only seven studies, all published after 2020 [44,60,66–70].

Number of posts included

For the publications which indicated it (n=57; 65.5%), the mean number of posts studied was 5,310, with a standard-deviation of 17,598, a minimum of 16 and a maximum of 120,738. The mean number of posts studied for publications using AI methods (n=7; 8.0%) was 33,201 compared to 1,405 for publications analysing the posts manually (n=50; 57.5%); the maximum is 120,738 compared to 13,089; and the minimum is 317 compared to 16.

Table 4. Number of posts included in the studies.

	Total	Studies with AI methods	Studies without AI methods
Number of studies giving the number of posts included	57	7	50
Mean	5,310	33,201	1,405
Standard deviation	17,598		
Minimum	16	317	16
Maximum	120,738	120,738	13,089

Temporal range

Among the publications which indicate it (n=42; 48.3%), almost 70% had a time range of less than three years (n=29; 33.3%). Obviously short temporal ranges were correlated with small datasets. Among the publications using AI methods, three had a temporal range of five years or more, accounting for half of the publications with a temporal range of five years or more.

Table 5. Number and proportion of publications by temporal range.

	Number of studies	Percentage of studies
[1 day; 1 year[	15	35.7
[1 year; 3 years[	14	33.3
[3 years; 5 years[	7	16.7
≥ 5 years	6	14.3

AI methods

Among the seven publications using AI methods, the two main AI-assisted tasks were sentiment analysis (n=2) [60,66] and topic identification (n=4) [44,66–68]. The other tasks tackled with AI methods were text generation (n=1) [69] and medical concepts identification with not much detail on the model used (n=1) [70]. Only three of them gave a score to evaluate the performance of the model: F1 scores [66], clarity and quality for the text generation task [69] and coherence [68].

For the topic identification task, the authors have implemented different approaches, from simple ones to machine learning methods. Chen et al. used regular expressions followed by manual analysis for topic identification [67]. On the other hand, Karas et al. [68] compared Latent Dirichlet

Allocation (LDA) with Top2Vec model [71] with six different embedding methods and obtained the best performance (Coherence = 0.642) with the Doc2vec embedding [72]. Bi et al. also used LDA on the comments of the posts they studied, and compared it with term frequency-inverse document frequency (TF-IDF) [66]. Each LDA topic was then manually annotated and a Naive Bayes classifier was trained on the comments to predict the topic [66]. Then, this classifier was used on the posts, in order to assign to each post one of the pre-identified topics (F1 score = 0.902) [66]. Another approach was Yao et al.'s [44], who used BERTopic [73].

Regarding the sentiment analysis task, Bi et al. compared a long-short term memory neural network (LSTM) to other classifiers such as Naive Bayes, Support Vector Machines, and convolutional neural networks [66]. With LSTM, they achieved a F1 score of 0.916. Bozkurt & Aras did not mention which model was used for the sentiment analysis task [60].

Finally, Schmälzle & Wilcox used existing tweets to generate new tweets, using medium-sized GPT-2 model [69]. Their objective was to assess the possibility of generating messages that appear natural to humans to promote awareness online [69]. They obtained a clarity of 3.58 for AI generated messages (compared to 3.34 for human generated messages) and a quality of 3.57 (compared to 3.3 for human generated messages) [69].

Discussion

Here we provide an overview of the studies published until 2023 related to rare diseases and using social media data. Our review aims at providing a "medical informatics" perspective on this field of research and to identify the main challenges to address the needs of the rare disease community. As such, this work aims at analysing the methods used for the "study of the determinants and distribution of health information" i.e., infodemiology [74]. Some limitations may be inherent to the selection and the analysis of the publications, although we used the methodology previously published by Miller et al. [15] and designed the protocol to be as objective as possible. In the following sections, we discuss the results brought by this review and the perspectives for the rare disease communities, the public health policy makers, and the artificial intelligence researchers.

Clinical significance

Rare diseases

Only four publications considered all rare diseases without focusing on a specific domain or disease. Another one, comparing social media activity in Sickle Cell Disease and Cystic Fibrosis was published in 2016 in a clinical journal. Overall, the other publications studied less than one hundred different rare diseases. Of note, this number represents less than 0.2% of the 5,000 to 8,000 existing rare diseases whereas the role of social media has been growing, globally and in the rare disease communities. This echoes the conclusion drawn by Miller et al. that "despite its potential benefits in rare disease research, the use of social media is still methodologically limited and the participants reached may not be representative of the rare disease population by gender, race, age, or rare disease type" [15]. We also claim that more rare diseases should benefit from this kind of research.

Patients and families

Almost half of the studies (44.8%) were based on messages posted by patients or their families. Among these, around 80% of the publications tackled many aspects of the rare disease such as quality of life, the needs expressed by patients and families, treatments and care pathways, etc. On the other hand, some (around 20%) addressed only one specific question such as the underuse of a given treatment, the genetic risk and tests, and very disease-specific aspects of patients' or caregivers' lives. The conclusion of these studies was that social media hold considerable potential to

a better understanding of patients' and caregivers' experiences. In the context of rare diseases, for which real-world data is limited, social media represent a valuable source of information to adapt care and treatment pathways. Moreover, social media provide patients and caregivers with a virtual space to inform themselves, express themselves and raise awareness.
Besides rare diseases per se, the impact of pandemics on rare disease communities is an important topic, related with the possible complications and the needs for follow-up and advice. This was demonstrated during the Covid-19 crisis [42–45].
Moreover, social media data may be used conjointly with other data such as surveys, interviews or scientific literature in order to define priorities or to provide models of natural histories [26,39,75].

Data science aspects

We partly reused the search methodology developed by Miller et al. [15] to explore the role of social media beyond clinical research and we extended the scope to more recent articles, and analysed them from a data science perspective.

AI methods

The use of AI to analyse social media is relatively new and still limited, but has increased in the last few years - with two studies in 2020/2021 and five in 2022/2023. Not surprisingly, we showed that the number of posts studied is on average higher with the use of AI methods (avg=33,201) than without them (avg=1,405). In terms of AI-tasks, sentiment analysis and topic detection were the main tasks. They are useful to grasp the overall sentiment of a message and to split messages in different groups respectively. However, they represent only a small part of what can currently be done with existing AI methods. We expect that more extensive use of AI could lead to major improvements, especially in the understanding of patients' and their families' experiences and needs.

Social media

The studies that we analysed used a wide range of existing social media. Although text messages remain the most important data source, videos and images are getting more attention. These trends are probably due to the rise of Instagram and TikTok, but also more generally to the rise of video formats on all social media platforms. Analysing videos and images, as well as understanding natural language data, still represents a major challenge. However, lots of current AI research efforts focus on these domains and recent progress has opened up new ways to mine social media. These advances will hopefully benefit rare disease research.

The majority of the studies (93.1%) were based on social media messages in English. Other languages were under-represented: only six non-English languages studied in our corpus, contrasting with the number of languages present on the Internet, and used by patients and their families all over the world. This may lead to difficulties in transposing the results from a study data set to another population. Several publications have mentioned possible biases in data sets associated with the usage of social media by patients and families with a lack of "representativeness of the broader rare disease community, both in terms of disease type and patient demographics" [15]. However, we believe that, especially for rare and ultra-rare diseases, we should extend our data coverage to non-English data. Ideally, the communities should promote access to more population-representative data in terms of countries and diseases to ensure accuracy for all populations.

Perspectives

This work allowed us to grasp what the reality of infodemiology for rare diseases research was in 2023. The opportunities of current AI research on NLP are still underexploited in this very specific field, resulting in an under exploitation of online data. AI based social media mining could entail a

much better understanding of rare disease patients' experiences and needs. Rare diseases are often chronic, progressive, degenerative, life-threatening and disabling diseases. Social media mining could be a way to investigate experience-based opinions in a quantitative way, and to complement other patient-driven initiatives like the surveys conducted by Eurordis [76]. Indeed, this approach has already shown promising results in other areas such as pandemics monitoring [77,78], adverse events monitoring [79,80], or quality of life evaluation [81].

This perspective is also in line with one of the three goals set by the International Rare Diseases Research Consortium (IRDiRC) for 2027 [82]. IRDiRC's working group identified that a preliminary selection of metrics highlighting how access to diagnostic and therapies impacts the health quality of rare disease patients, the socio-economic burden on patients and families, and the economy and efficiency of HCS and insurance companies was needed [4]. They also showed that the most important factors to consider according to patients and families are quality of life or health outcomes and the socio-economic burden of rare diseases [4]. To measure the socio-economic burden of rare diseases (and of diagnosis and therapies), they identified data elements that could be grouped into four broad categories: diagnostics, prevalence, natural history studies and intervention [4]. Moreover, they recommended collecting real-world evidence data for the natural history studies [4]. Harnessing online patient or caregiver-generated data could be one way to go.

Contrasting with such high expectancies, this review shows that social media based studies in the rare disease field are still at an early stage, with only a tiny portion of rare diseases studied, with only a few languages studied also, and mainly with only very few studies exploiting current NLP progress to extract knowledge from social media data.

Conclusion

Social media mining in the rare disease area is still a research domain that has not benefited from recents advances in AI. The use of social media user-generated data could provide patients, health professionals, and researchers with information that is not accessible otherwise. It could also help tackle the issues arising from the low prevalence of the diseases, the need for support and the possible isolation of the patients. In other words, real-world data from social media should play an important role in rare disease research. Our results suggest that it could be increasingly developed and should employ innovative methods and approaches.

Acknowledgements

This work was supported by state funding by The French National Research Agency (ANR) as part of the "Investissements d'avenir" program (ANR-19-P3IA-0001) (PaRis AI Research InstitutE (PR [AI]RIE)).

Conflicts of interest

None declared

Abbreviations

AI: Artificial Intelligence
CLP: Cleft Lip/Palate
EURORDIS: European Organisation for Rare Diseases
HU: Hydroxyurea
IRDiRC: International Rare Diseases Research Consortium
LDA: Latent Dirichlet Allocation
LSTM: long-short term memory neural network

MND: Motor Neuron Diseas
NLP: Natural Language Processing
SCD: Sickle Cell Disease
TF-IDF: term frequency-inverse document frequency

Multimedia Appendix 1

Excel file with all included publications.

References

1. Baekelandt E. Background Paper 6.19 Rare Diseases. 2013. Available from: https://www.semanticscholar.org/paper/Background-Paper-6.19-Rare-Diseases-Baekelandt/6c6fd13aab279bdee436b4725a4408d81bf8d39b [accessed Jun 13, 2023]
2. The long journey to a rare disease diagnosis | Research and Innovation. Available from: https://ec.europa.eu/research-and-innovation/en/horizon-magazine/long-journey-rare-disease-diagnosis [accessed Aug 11, 2023]
3. Benito-Lozano J, Arias-Merino G, Gómez-Martínez M, Arconada-López B, Ruiz-García B, Posada de la Paz M, Alonso-Ferreira V. Psychosocial impact at the time of a rare disease diagnosis. Fernandez-Lozano C, editor. PLOS ONE 2023 Jul 28;18(7):e0288875. doi: 10.1371/journal.pone.0288875
4. Zanello G, Chan C-H, Pearce DA, IRDiRC Working Group. Recommendations from the IRDiRC Working Group on methodologies to assess the impact of diagnoses and therapies on rare disease patients. Orphanet J Rare Dis 2022 Dec;17(1):181. doi: 10.1186/s13023-022-02337-2
5. Matthews L, Chin V, Taliangis M, Samanek A, Baynam G. Childhood rare diseases and the UN convention on the rights of the child. Orphanet J Rare Dis 2021 Dec;16(1):523. doi: 10.1186/s13023-021-02153-0
6. Mukherjee K. Care for Rare: Spotlight on Rare Diseases. Trends Pharmacol Sci 2019 Apr;40(4):227–228. doi: 10.1016/j.tips.2019.02.008
7. Nguengang Wakap S, Lambert DM, Olry A, Rodwell C, Gueydan C, Lanneau V, Murphy D, Le Cam Y, Rath A. Estimating cumulative point prevalence of rare diseases: analysis of the Orphanet database. Eur J Hum Genet 2020 Feb;28(2):165–173. doi: 10.1038/s41431-019-0508-0
8. Yang G, Cintina I, Pariser A, Oehrlein E, Sullivan J, Kennedy A. The national economic burden of rare disease in the United States in 2019. Orphanet J Rare Dis 2022 Dec;17(1):163. doi: 10.1186/s13023-022-02299-5
9. Velvin G, Dammann B, Haagensen T, Johansen H, Strømme H, Geirdal AØ, Bathen T. Work participation in adults with rare genetic diseases - a scoping review. BMC Public Health 2023 May 19;23(1):910. doi: 10.1186/s12889-023-15654-3
10. Delaye J, Cacciatore P, Kole A. Valuing the “Burden” and Impact of Rare Diseases: A Scoping Review. Front Pharmacol 2022 Jun 8;13:914338. doi: 10.3389/fphar.2022.914338
11. Juggling care and daily life: The balancing act of the rare disease community. EURORDIS. Available from: https://www.eurordis.org/publications/juggling-care-and-daily-life-the-balancing-act-of-the-rare-disease-community/ [accessed Aug 11, 2023]
12. Mombini H, Li R, Zhang Y, Korkin D, Tulu B. [3] An Exploratory Study of Social Media Analysis for Rare Diseases Using Machine Learning Algorithms: A Case Study of Trigeminal Neuralgia. 2020. doi: 10.24251/HICSS.2020.469
13. World Bank Open Data. World Bank Open Data. Available from: https://data.worldbank.org [accessed Aug 11, 2023]
14. Digital 2022: Global Overview Report. DataReportal – Glob Digit Insights. 2022. Available

from: https://datareportal.com/reports/digital-2022-global-overview-report [accessed Aug 11, 2023]
15. Miller EG, Woodward AL, Flinchum G, Young JL, Tabor HK, Halley MC. Opportunities and pitfalls of social media research in rare genetic diseases: a systematic review. Genet Med 2021 Dec;23(12):2250–2259. doi: 10.1038/s41436-021-01273-z
16. Pérez Dasilva J, Santos Diez MT, Meso Ayerdi K. Las asociaciones de enfermedades raras: Estructura de sus redes e identificación de los líderes de opinión mediante la técnica del análisis de redes sociales. Rev Lat 2021 Apr 7;(79):175–205. doi: 10.4185/RLCS-2021-1498
17. Sánchez-Castillo S, Mercado-Sáez M-T. Sufro una grave enfermedad rara. Reto a cantar y hacer coreografías en TikTok. El Prof Inf 2021 Jul 27;e300414. doi: 10.3145/epi.2021.jul.14
18. Subirats L, Reguera N, Bañón AM, Gómez-Zúñiga B, Minguillón J, Armayones M. Mining Facebook Data of People with Rare Diseases: A Content-Based and Temporal Analysis. Int J Environ Res Public Health Multidisciplinary Digital Publishing Institute; 2018 Sep;15(9):1877. doi: 10.3390/ijerph15091877
19. Titgemeyer SC, Schaaf CP. Facebook Support Groups for Rare Pediatric Diseases: Quantitative Analysis. JMIR Pediatr Parent 2020 Nov 19;3(2):e21694. doi: 10.2196/21694
20. Barriteau CM, Thompson AL, Meier ER, Pecker LH. Sickle cell disease related internet activity is three times less frequent than cystic fibrosis related internet activity. Pediatr Blood Cancer 2016 Nov;63(11):2061–2062. PMID:27362449
21. Korkmaz YN, Arslan S, Buyuk SK. Bullying in individuals with cleft lip and palate: A Twitter analysis. Int J Clin Pract 2021 Nov;75(11). doi: 10.1111/ijcp.14856
22. Albright K, Walker T, Baird S, Eres L, Farnsworth T, Fier K, Kervitsky D, Korn M, Lederer DJ, McCormick M, Steiner JF, Vierzba T, Wamboldt FS, Swigris JJ. Seeking and sharing: why the pulmonary fibrosis community engages the web 2.0 environment. BMC Pulm Med 2016 Jan 12;16:4. PMID:26754048
23. Black R, Freifeld C, Hogue S, Decoteau S, Pham S. PCR102 Leveraging Social Media for Patient Experience Insights in Rare Disease. Value Health 2022 Jul 1;25(7, Supplement):S560. doi: 10.1016/j.jval.2022.04.1445
24. Gajjar A, Jain A, Le AH-D, Salem MM, Jankowitz BT, Burkhardt J-K. Cerebral Cavernous Malformations Patient Perception Analysis via Social Media. J Neurol Surg Part Cent Eur Neurosurg 2022 Dec 8;a-1994-9435. doi: 10.1055/a-1994-9435
25. Gajjar A, Jain A, Salem M, Yueh Hou N, Huy Dinh A, Jankowitz B, Burkhardt J. E-209 Patients' perceptions of moyamoya disease on instagram and tiktok. SNIS 19th Annu Meet Electron Poster Abstr BMJ Publishing Group Ltd.; 2022. p. A192.1-A192. doi: 10.1136/neurintsurg-2022-SNIS.320
26. Goodspeed K, Mosca LR, Weitzel NC, Horning K, Simon EW, Pfalzer AC, Xia M, Langer K, Freed A, Bone M, Picone M, Bichell TJV. A draft conceptual model of SLC6A1 neurodevelopmental disorder. Front Neurosci 2023 Jan 19;16:1026065. PMID:36741059
27. Liao BT, Busse J, Ender KL, Schechter WS. Exploring social media for patient perspectives of sickle cell disease. Pediatr Hematol Oncol 2016 Mar;33(2):134–135. PMID:26934042
28. Maurya A, Dixit G, George A, Karki R, Aasaithambi S, Verma H, Doherty J. POSB352 Mining Socialmedia to Understand Unmet Need and Treatment Experience in Patients with Atypical Hemolytic Uremic Syndrome (AHUS). Value Health 2022 Jan;25(1):S229. doi: 10.1016/j.jval.2021.11.1119
29. Spies E, Andreu T, Koelling J, Hartung M, Kamudoni P, Park J. PCR18 An Exploratory Retrospective Social Listening Study to Identify Patient Experiences Associated With Cutaneous Lupus Erythematosus (CLE). Value Health 2022 Dec;25(12):S393. doi: 10.1016/j.jval.2022.09.1953
30. Strobel MJ, Alves D, Roufosse F, Antoun Z, Kwon N, Baylis L, Wechsler ME. Insights from Social Media on the Patient Experience of Living With Rare Eosinophil-Driven Diseases. J

Patient Exp 2022;9:23743735221143950. PMID:36530646
31. Umbaugh HM, Crerand CE, Stock NM, Luquetti DV, Heike CL, Drake AF, Billaud Feragen KJ, Johns AL. Microtia and craniofacial microsomia: Content analysis of facebook groups. Int J Pediatr Otorhinolaryngol 2020 Nov 1;138:110301. doi: 10.1016/j.ijporl.2020.110301
32. Wang T, Lund B. Categories of Information Need Expressed by Parents of Individuals with Rare Genetic Disorders in a Facebook Community Group: A Case Study with Implications for Information Professionals. J Consum Health Internet Routledge; 2020 Jan 2;24(1):20–34. doi: 10.1080/15398285.2020.1713700
33. Stewart A, Smith CH, Govender R, Eaton S, De Coppi P, Wray J. Parents' experiences of feeding children born with oesophageal atresia/tracheo-oesophageal fistula. J Pediatr Surg 2022 Dec;57(12):792–799. PMID:36150934
34. Walker AL, Gaydos LM, Farzan R, De Castro L, Jonassaint C. Social media discussions provide new insight about perceptions of hydroxyurea in the sickle cell community. Am J Hematol 2019;94(5):E134–E136. doi: 10.1002/ajh.25430
35. Walker KK. A content analysis of cognitive and affective uses of patient support groups for rare and uncommon vascular diseases: comparisons of may thurner, thoracic outlet, and superior mesenteric artery syndrome. Health Commun 2015;30(9):859–871. PMID:24877701
36. Kline E, Garrett AL, Brownstein C, Ziniel S, Payton E, Goldin A, Hoffman K, Chandler J, Weber S. Using social media listening to understand barriers to genomic medicine for those living with Ehlers-Danlos syndromes and hypermobility spectrum disorders. Health Expect Int J Public Particip Health Care Health Policy 2023 Apr 16; PMID:37062887
37. Cotton AC, Bell RB, Jinnah HA. Expert Opinion vs Patient Perspective in Treatment of Rare Disorders: Tooth Removal in Lesch-Nyhan Disease as an Example. JIMD Rep 2017 Dec 15;41:25–27. PMID:29243037
38. Nakamura C, Bromberg M, Bhargava S, Wicks P, Zeng-Treitler Q. Mining Online Social Network Data for Biomedical Research: A Comparison of Clinicians' and Patients' Perceptions About Amyotrophic Lateral Sclerosis Treatments. J Med Internet Res 2012 Jun 21;14(3):e90. PMID:22721865
39. Mahoney AF, Handberg C. New medicine for neuromuscular diseases: An evolving paradox for patient and family hopes and expectations. Nurs Inq 2023 Apr;30(2):e12527. doi: 10.1111/nin.12527
40. Howard J, Mazanderani F, Locock L. Life 'on high alert': how do people with a family history of motor neurone disease make sense of genetic risk? insights from an online forum. Health Risk Soc 2021 Aug 18;23(5–6):179–195. doi: 10.1080/13698575.2021.1946488
41. Smedley RM, Coulson NS. Genetic testing for Huntington's disease: A thematic analysis of online support community messages. J Health Psychol 2021 Mar;26(4):580–594. doi: 10.1177/1359105319826340
42. Reuter K, Deodhar A, Makri S, Zimmer M, Berenbaum F, Nikiphorou E. The impact of the COVID-19 pandemic on people with rheumatic and musculoskeletal diseases: insights from patient-generated data on social media. Rheumatology 2021 Oct 9;60(SI):SI77–SI84. doi: 10.1093/rheumatology/keab174
43. Wray J, Pagel C, Chester AH, Kennedy F, Crowe S. What was the impact of the first wave of COVID-19 on the delivery of care to children and adults with congenital heart disease? A qualitative study using online forums. BMJ Open 2021 Sep;11(9):e049006. doi: 10.1136/bmjopen-2021-049006
44. Yao LF, Ferawati K, Liew K, Wakamiya S, Aramaki E. Disruptions in the Cystic Fibrosis Community's Experiences and Concerns During the COVID-19 Pandemic: Topic Modeling and Time Series Analysis of Reddit Comments. J Med Internet Res 2023 Apr 20;25:e45249. PMID:37079359
45. Stewart A, Smith CH, Eaton S, De Coppi P, Wray J. COVID-19 pandemic experiences of

parents caring for children with oesophageal atresia/tracheo-oesophageal fistula. BMJ Paediatr Open 2021 May;5(1):e001077. doi: 10.1136/bmjpo-2021-001077
46. Coulson NS, Buchanan H, Aubeeluck A. Social support in cyberspace: A content analysis of communication within a Huntington's disease online support group. Patient Educ Couns 2007 Oct 1;68(2):173–178. doi: 10.1016/j.pec.2007.06.002
47. Wittmeier K, Holland C, Hobbs-Murison K, Crawford E, Beauchamp C, Milne B, Morris M, Keijzer R. Analysis of a parent-initiated social media campaign for Hirschsprung's disease. J Med Internet Res 2014 Dec 11;16(12):e288. PMID:25499427
48. Slick N, Bodas P, Badawy SM, Wildman B. Accuracy of online medical information: the case of social media in sickle cell disease. Pediatr Hematol Oncol 2023 Mar;40(2):99–107. PMID:35635234
49. Jones P, Rajasegaran A, Brassale S, Chen Y, Haslam R, Austin C, Seideman CA. Assessment of the Educational Value of Distal Hypospadias Repair Videos on YouTube. Urology 2022 Jan;159:28–32. PMID:34461144
50. Karabay E, Karsiyakali N, Kayar K, Koseoglu H. Hypospadias surgery on YouTube: is it valid? Minerva Pediatr 2021 May;73(3). doi: 10.23736/S2724-5276.19.05555-5
51. Karakoyun A, Yildirim A. YouTube videos as a source of information concerning Behçet's disease: a reliability and quality analysis. Rheumatol Int 2021 Dec;41(12):2117–2123. doi: 10.1007/s00296-021-05009-9
52. Krakowiak M, Szmuda T, Fercho J, Ali S, Maliszewska Z, Słoniewski P. YouTube as a source of information for arteriovenous malformations: A content-quality and optimization analysis. Clin Neurol Neurosurg 2021 Aug;207:106723. doi: 10.1016/j.clineuro.2021.106723
53. Küçükakkaş O, İnce B. Can YouTube be used as an educational tool in lymphedema rehabilitation? Arch Physiother 2022 Mar 3;12(1):5. PMID:35236412
54. Sasse M, Ohrndorf S, Palmowski A, Wagner AD, Burmester GR, Pankow A, Krusche M. POS0359-PARE YouTube AS A SOURCE OF INFORMATION ON AUTOINFLAMMATORY DISEASES. Ann Rheum Dis 2022 Jun;81(Suppl 1):432.1-433. doi: 10.1136/annrheumdis-2022-eular.4016
55. Sasse M, Ohrndorf S, Palmowski A, Wagner AD, Burmester GR, Pankow A, Krusche M. Digital health information on autoinflammatory diseases: a YouTube quality analysis. Rheumatol Int 2023 Jan;43(1):163–171. PMID:36374326
56. Srivastav S, Tewari N, Antonarakis GS, Upadhyaya AD, Duggal R, Goel S. How Informative Is YouTube Regarding Feeding in Infants with Cleft Lip and Palate? Cleft Palate Craniofacial J 2022 Dec 14;105566562211421. doi: 10.1177/10556656221142194
57. Wilkens F, Ganter C, Kriegsmann K, Wilkens H, Kahn N, Goobie GC, Ryerson CJ, Kreuter M. Is YouTube a reliable source for patient information in lymphangioleiomyomatosis? Eur Respir J European Respiratory Society; 2022 Sep 4;60(suppl 66). doi: 10.1183/13993003.congress-2022.85
58. Castillo J, Wassef C, Wassef A, Stormes K, Berry AE. YouTube as a Source of Patient Information for Prenatal Repair of Myelomeningocele. Am J Perinatol 2021 Jan;38(02):140–144. doi: 10.1055/s-0039-1694786
59. Henrick V, Marks S, Balmer R, Barber S. Public and dental professionals' use of social media to discuss amelogenesis imperfecta. Int J Paediatr Dent 2022 Nov;32(6):903–914. PMID:35771161
60. Bozkurt AP, Aras I. Cleft Lip and Palate YouTube Videos: Content Usefulness and Sentiment Analysis. Cleft Palate Craniofac J 2021 Mar;58(3):362–368. doi: 10.1177/1055665620948722
61. Guthrie G, Davies RM, Fleming CK, Browning AC. YouTube as a source of information about retinitis pigmentosa. Eye 2014 Apr;28(4):499–500. PMID:24434660
62. Charbonneaux J, Berthelot-Guiet K. Portraits of a Killer: Visual Expressions of Patient/Parent Expertise on Social Media. In: Meiselwitz G, Moallem A, Zaphiris P, Ioannou A, Sottilare RA,

Schwarz J, Fang X, editors. HCI Int 2022 - Late Break Pap Interact New Media Learn Games Cham: Springer Nature Switzerland; 2022. p. 30–40. doi: 10.1007/978-3-031-22131-6_3
63. Frantz LM, Wall LB, Goldfarb CA. Media Depiction of Birth Differences of the Upper Extremity: Accuracy of Shared Diagnoses. J Pediatr Orthop 2022 Aug;42(7):e753–e755. doi: 10.1097/BPO.0000000000002185
64. Kilcoyne S, Rogers C, Thomas GPL, Wall S, Johnson D. Craniofacial Surgery-Related Hashtag Utilisation on Instagram. J Craniofac Surg 2021 Sep;32(6):2035–2040. doi: 10.1097/SCS.0000000000007593
65. Parker C, Zomer E, Liew D, Ayton D. Characterising experiences with acute myeloid leukaemia using an Instagram content analysis. Laws MB, editor. PLOS ONE 2021 May 3;16(5):e0250641. doi: 10.1371/journal.pone.0250641
66. Bi Q, Shen L, Evans R, Zhang Z, Wang S, Dai W, Liu C. Determining the Topic Evolution and Sentiment Polarity for Albinism in a Chinese Online Health Community: Machine Learning and Social Network Analysis. JMIR Med Inform 2020 May 29;8(5):e17813. PMID:32469320
67. Chen R, Muralidharan K, Samelson-Jones BJ. Digital haemophilia: Insights into the use of social media for haemophilia care, research and advocacy. Haemoph Off J World Fed Hemoph 2022 Mar;28(2):247–253. PMID:35167716
68. Karas B, Qu S, Xu Y, Zhu Q. Experiments with LDA and Top2Vec for embedded topic discovery on social media data-A case study of cystic fibrosis. Front Artif Intell 2022;5:948313. PMID:36062265
69. Schmälzle R, Wilcox S. Harnessing Artificial Intelligence for Health Message Generation: The Folic Acid Message Engine. J Med Internet Res 2022 Jan 18;24(1):e28858. PMID:35040800
70. Picone M, Ascencion F, Wassman ER, DeFelice C, Hernandez HW, Converse M, Flowers C, Flurie M, Horsnell M. Extracting Real-World Insights From Social Media to Understand Narcolepsy and the Impact of Brain Fog. Sleep Med 2022 Dec;100:S159. doi: 10.1016/j.sleep.2022.05.431
71. Angelov D. Top2Vec: Distributed Representations of Topics. arXiv; 2020; doi: 10.48550/ARXIV.2008.09470
72. Le QV, Mikolov T. Distributed Representations of Sentences and Documents. arXiv; 2014; doi: 10.48550/ARXIV.1405.4053
73. Grootendorst M. BERTopic: Neural topic modeling with a class-based TF-IDF procedure. arXiv; 2022. doi: 10.48550/arXiv.2203.05794
74. Eysenbach G. Infodemiology: the epidemiology of (mis)information. Am J Med 2002 Dec;113(9):763–765. doi: 10.1016/S0002-9343(02)01473-0
75. Ardouin K, Davis N, Stock NM. Expanding Support Services for Adults Born With Cleft Lip and/or Palate in the United Kingdom: An Exploratory Evaluation of the Cleft Lip and Palate Association Adult Services Programme. Cleft Palate-Craniofacial J Off Publ Am Cleft Palate-Craniofacial Assoc 2022 Apr;59(4_suppl2):S48–S56. PMID:34184577
76. The Voice of 12,000 Patients. EURORDIS. Available from: https://www.eurordis.org/publications/the-voice-of-12000-patients/ [accessed Aug 21, 2023]
77. Déguilhem A, Malaab J, Talmatkadi M, Renner S, Foulquié P, Fagherazzi G, Loussikian P, Marty T, Mebarki A, Texier N, Schuck S. Identifying Profiles and Symptoms of Patients With Long COVID in France: Data Mining Infodemiology Study Based on Social Media. JMIR Infodemiology 2022 Nov 22;2(2):e39849. doi: 10.2196/39849
78. Schück S, Foulquié P, Mebarki A, Faviez C, Khadhar M, Texier N, Katsahian S, Burgun A, Chen X. Concerns Discussed on Chinese and French Social Media During the COVID-19 Lockdown: Comparative Infodemiology Study Based on Topic Modeling. JMIR Form Res 2021 Apr 5;5(4):e23593. doi: 10.2196/23593
79. Arnoux-Guenegou A, Girardeau Y, Chen X, Deldossi M, Aboukhamis R, Faviez C, Dahamna B, Karapetiantz P, Guillemin-Lanne S, Louët AL-L, Texier N, Burgun A, Katsahian S. The

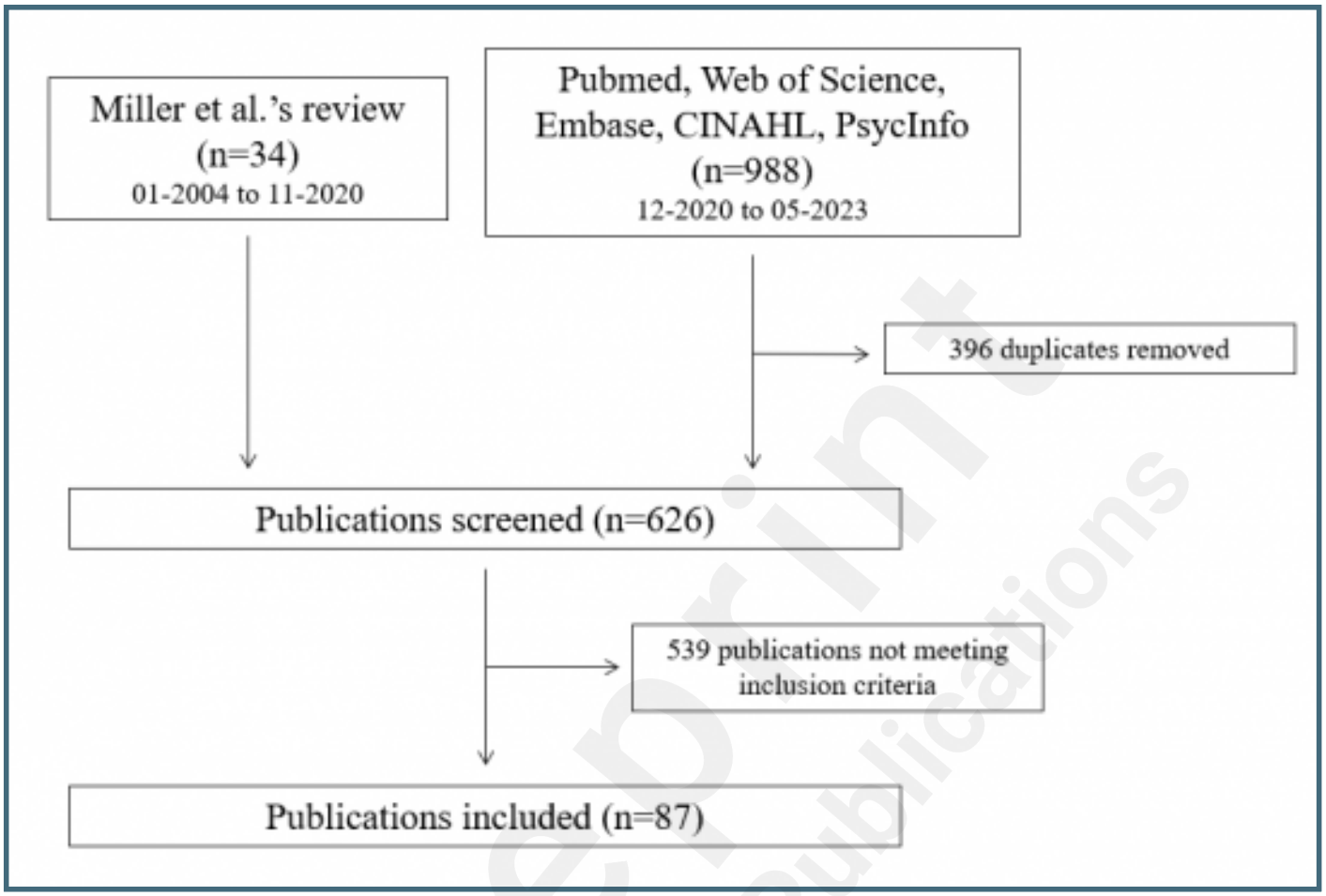
Adverse Drug Reactions From Patient Reports in Social Media Project: Protocol for an Evaluation Against a Gold Standard. JMIR Res Protoc 2019 May 7;8(5):e11448. doi: 10.2196/11448

80. Chen X, Faviez C, Schuck S, Lillo-Le-Louët A, Texier N, Dahamna B, Huot C, Foulquié P, Pereira S, Leroux V, Karapetiantz P, Guenegou-Arnoux A, Katsahian S, Bousquet C, Burgun A. Mining Patients' Narratives in Social Media for Pharmacovigilance: Adverse Effects and Misuse of Methylphenidate. Front Pharmacol 2018;9. Available from: https://www.frontiersin.org/articles/10.3389/fphar.2018.00541 [accessed Aug 23, 2023]
81. Eteve-Pitsaer C, Marty T, Nguyen A, Le Priol E, Paris C, Mebarki A, Texier N, Schück S. Psoriasis et altérations de la qualité de vie au travail: une étude avec des données issues de la base THIN® France croisées avec les contenus des réseaux sociaux analysés par l'outil Detec't®. Rev DÉpidémiologie Santé Publique 2022 Nov 1;70:S281–S282. doi: 10.1016/j.respe.2022.09.015
82. Vision & Goals – IRDiRC. Available from: https://irdirc.org/about-us/vision-goals/ [accessed Aug 15, 2023]

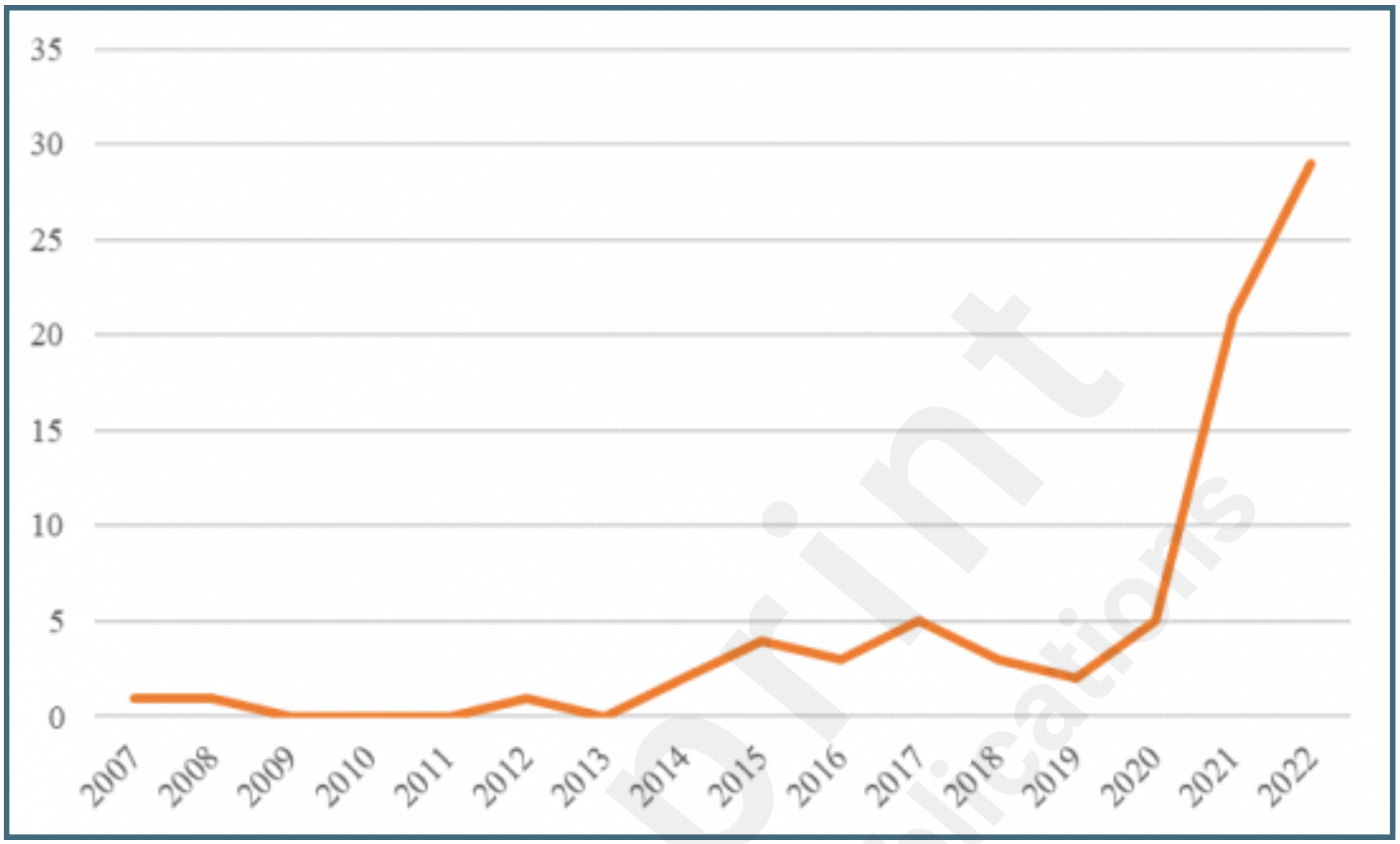
Supplementary Files

Figures

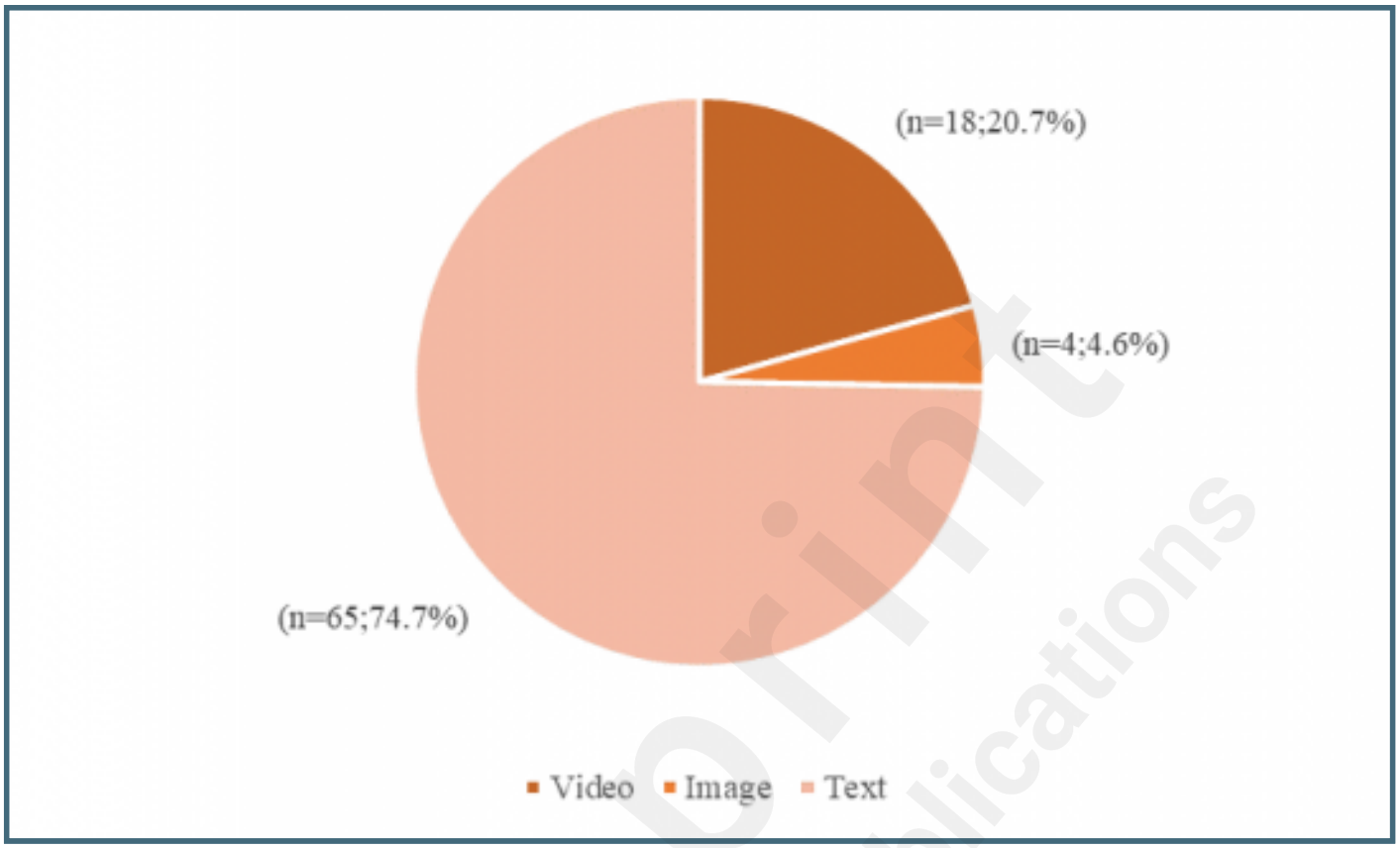
Flow diagram.



Number of included publications by year of publication.



Type of content analysed.



Multimedia Appendixes

Untitled.
URL: http://asset.jmir.pub/assets/80d24c3dd4b5533f749e6fbad369a0d3.xlsx