PhD-FSTM-2023-144
The Faculty of Science, Technology and Medicine

DISSERTATION

Defence held on 05/12/2023 in Luxembourg

to obtain the degree of

DOCTEUR DE L'UNIVERSITÉ DU LUXEMBOURG

EN BIOLOGIE

by

Charline Bour

Born on 20 April 1998 in Thionville, (France)

Use of artificial intelligence methods for the analysis of real-world and social media data in digital epidemiology

Dissertation defence committee

Dr Guy Fagherazzi, dissertation supervisor
Director of Department of Precision Health,
Integrated Group Leader Deep Digital Phenotyping Research Unit, Luxembourg Institute of Health

Dr Enrico Glaab, Chairman
Assistant Professor, University of Luxembourg

Dr Adam Hulman, Member
Associate Professor in Predictive Modelling, Department of Public Health, Aarhus University

Dr Petr Nazarov, Member
Group Leader, Multiomics Data Science, Head of Bioinformatics Platform, Luxembourg Institute of Health

Dr. Pierre-Yves Benhamou, Vice-Chair
MD, Endocrinologist, CHU Grenoble Alpes, France

Table of content

Acknowledgements

For the last three years, I've navigated the tumultuous seas of research and discovery, like a pirate on a quest for treasure, with the red cross on the map representing the final defence. Throughout this adventure, even when the winds were favorable and the tides on my side, the greatest challenge often came from within, battling against the self-doubt and the internal critique. Still, in such battles, I was lucky enough to be surrounded by an amazing crew -my mentor, peers, friends, and loved ones- who fought for me and by my side.

I would like to start by thanking my supervisor, **Dr. Guy Fagherazzi**, Director of Department of Precision Health, Group Leader Deep Digital Phenotyping Research Unit. I want to thank you for giving me the chance to join your team and explore a field that was once new to me, and in which I now thrive. Working with you made being an outstanding supervisor look easy. Stories of insufficient guidance are all too common in thesis work, but I was lucky to always have your support. I'm deeply thankful for that.

I would like to extend my gratitude to **Dr. Enrico Glaab** and **Dr. Maria Ruiz** for being part of my thesis monitoring committee and for your invaluable advices. I also thank you for accepting to assess my work and to serve on the thesis defence commitee along with **Dr. Petr Nazarov**, **Prof. Pierre-Yves Benhamou** and **Dr. Adam Hulman**.

Thank you, thank you, thank you, and a million times over, to every current and former member of the Deep Digital Phenotyping research team. Working with colleagues like you, who also happen to often bring cakes to share, has been an incredible motivator to come to work every day! **Mégane** and **Maurane**, thank you for sharing in both my joys, frustrations, and for the pictures of cute cats/funny memes that always brightened my days. **Abir**, my work bestie, you're fun, talented, and an endless source of support, thank you for always being there. I wish everyone to have a Abir in their life. **Kevser, Hanin, Dulce**, your kind words, smiles and attentive ears were always there, thank you. **Aurélie** and **Gloria,** you were always there with good advice or a helping hand when I needed it. Thank you so much.

I had so much fun thanks to all of you. You are all extraordinary just the way you are. Never change.

I also extend my thanks to **Aurore**, **Catherine, Franchesca**, **Luigi**, and everyone else at LIH and the University of Luxembourg who gave their time to guide, advise and assist me throughout this thesis journey.

Many thanks to my friends, and particularly my bébous, **Manon** and **Valentin**. Whether it was over a beer (or a manzanita), a board game, or video games—sometimes all three at once—I've enjoyed incredible moments that allowed me to step back from my thesis work and recharge.

Nolwen, your support even from the other end of France, provided comfort and encouragement when times were tough, thank you.

A huge thank you to my bébou #1, **Emeric**, who has constantly pushed me to excel and give my best. If I have reached this point today, it is largely thanks to you and your influence. I began these acknowledgments with an analogy to a pirate's journey, and if I am the pirate, you are the lighthouse, always guiding me towards happiness and safeness. Thank you.

I would also like to express my gratitude to my family for their support. You all laughed at the acknowledgments in my master's thesis, finding humor in the ordinary that in your eyes is just another day of being the incredible family you are. I'm curious to see your reaction to these ones. **Christophe**, my brother, for a decade you've jokingly called me 'doctor,' and now I can proudly claim that title for real. **Mum**, **Dad**, and **Christophe**, your constant presence, the sacrifices you've made, and the way you celebrate each of my accomplishments may seem ordinary to you, but they mean the world to me. You think you've done nothing special, but that couldn't be further from the truth. Thank you for everything; I am truly fortunate to be surrounded by a family like ours.

Abbreviations

ALTRUIST: virtuAL digiTal cohoRt stUdy on TwItter uSing pyThon

AI: Artificial Intelligence

API: Application Programming Interface

BERT**:** Bidirectional Encoder Representations from Transformers

LIWC: Linguistic Inquiry and Word Count

ML: Machine Learning

PWD: People with Diabetes

SM: Social Media

VDCS: Virtual Digital Cohort Study

WDDS: World Diabetes Distress Study

Abstract

Introduction

Among all the digital data sources, social media have emerged as a significant source of health-related information, offering access to patient perspectives, outcomes and experiences. This rapid access to patients' emotions and concerns at a large scale represents a unique opportunity to improve patient-centered research and care. Relying on its dynamic online community, this thesis will focus on people living with diabetes, with an aim to better describe and understand the burden of diabetes. As part of the World Diabetes Distress Study, this thesis explores the potential of using social media data for health research and digital epidemiology using artificial intelligence methods for chronic diseases to go beyond the historical analysis of online data to monitor infectious disease epidemics. The overall aim is to demonstrate how social media data can capture key insights from health-related discussions and shape and enhance healthcare strategies.

Methods

We first used a scoping review approach to identify all the different uses of social media for health research purposes. Second, a global analysis of diabetes-related tweets was conducted to identify the critical determinants of diabetes burden and the differences in how diabetes is perceived worldwide. Then, we developed the concept of a virtual digital cohort study (VDCS) and designed a specialized tool to standardize and analyze social media data as a typical cohort study.

Results

We have shown that social media platforms can be used for health research. It can be used for various tasks, from recruitment to the dissemination of information and data collection. The rich information shared by the communities of people with diabetes can be used as a complementary approach to traditional, questionnaire-based epidemiology. This project led to the analysis of 54 million diabetes-related tweets collected between 2017 and 2021, thereby enhancing our understanding of the diabetes burden worldwide. An open-source Python package ALTRUIST was created to standardize and simplify setting up VDCS using social media data. It allows researchers to effectively navigate through various stages of data collection, pre-processing, and analysis, simulating a traditional cohort study using social media data.

Discussion/Conclusion

This research highlights the potential of social media data in health research and digital epidemiology. Social media data can give a valuable, unbiased, unrestricted and unfiltered insight into patients' daily lives and experiences complementary to traditional approaches. The ALTRUIST package was designed to standardize the analysis of such data like a cohort and to help the research community to develop social-media-based research projects. Significant ethical, methodological and technical challenges remain to be addressed as we continue to deepen the field. Therefore, standardization of methodologies is necessary to gain the impact of results and trust from healthcare professionals. This work can be considered a first step towards a cohesive, standardized field to boost patient-centered care and global health strategies.

List of tables and figures

Figures

Tables

List of publications

Publications that are part of the thesis

1. **Bour C**, Ahne A, Schmitz S, Perchoux C, Dessenne C, Fagherazzi G. The Use of Social Media for Health Research Purposes: Scoping Review. J Med Internet Res. 2021 May 27;23(5):e25736. doi: 10.2196/25736. PMID: 34042593; PMCID: PMC8193478.
2. **Bour C,** Ahne A, Aguayo G, et al, Global diabetes burden: analysis of regional differences to improve diabetes care BMJ Open Diabetes Research and Care 2022;10:e003040. doi: 10.1136/bmjdrc-2022-003040
3. Bour C, Elbeji A, De Giovanni L, Ahne A, Fagherazzi G, ALTRUIST: a Python package to emulate a Virtual Digital Cohort Study on Twitter, Submitted to IEEE Transactions on Big Data

Other publications

1. **Bour C**, Schmitz S, Ahne A, et alScoping review protocol on the use of social media for health research purposes, BMJ Open 2021;11:e040671. doi: 10.1136/bmjopen-2020-040671
2. Ahne, A., Khetan, V., Tannier, X., Rizvi, M.I.H., Czernichow, T., Orchard, F., **Bour, C.**, Fano, A. and Fagherazzi, G., 2021. Identifying causal associations in tweets using deep learning: Use case on diabetes-related tweets from 2017-2021. arXiv preprint arXiv:2111.01225.
3. Ahne A, Khetan V, Tannier X, Rizvi M, Czernichow T, Orchard F, **Bour C**, Fano A, Fagherazzi G, Extraction of Explicit and Implicit Cause-Effect Relationships in Patient-Reported Diabetes-Related Tweets From 2017 to 2021: Deep Learning Approach, JMIR Med Inform 2022;10(7):e37201, DOI: 10.2196/37201
4. Fagherazzi, G., **Bour, C.**, & Ahne, A. (2022). Emulating a virtual digital cohort study based on social media data as a complementary approach to traditional epidemiology: when, what for, and how?. Diabetes Epidemiology and Management, 100085.
5. Ayadi, H., **Bour, C.**, Fischer, A., Ghoniem, M., & Fagherazzi, G. The Long COVID experience from a patient's perspective: a clustering analysis of 27,216 Reddit posts. Frontiers in Public Health, 11, 1227807.

Chapter 1

Objectives

This thesis is part of the World Diabetes Distress Study (WDDS) project funded by MSD Avenir that aims to study the complex interplays between lifestyle, psychological factors, and biological parameters that are key in the daily management of type 1 or type 2 diabetes and the development of diabetes-related complications. In traditional epidemiology, such a study would be both costly and time-consuming.

Digital epidemiology focuses on studying, for health-research purposes, data from the digitosome, an emerging concept gathering all the data generated online by an individual, including social media. Social media represents a unique resource that can be analyzed using big data and artificial intelligence.

To achieve the objective of WDDS, the thesis has the following specific aims:

1) To study the different uses of social media for health research, along with the various features, methodologies, and guidelines available for researchers.
2) To assess and identify the determinants of diabetes burden from the perspective of people with diabetes worldwide, using artificial intelligence methods to provide insights into the cultural and environmental factors that affect the daily management of diabetes.
3) To develop and validate a methodology for replicating traditional cohort studies with social media data, using machine learning (ML) and natural language processing techniques, to analyze large-scale and longitudinal data on lifestyle and psychological factors in a complementary, cost-effective and efficient manner.

Chapter 2

Synopsis

Traditional to digital epidemiology

The origins

The word “Epidemiology” comes from the Greek word “Epidemios”, meaning “among the people of one's countrymen at home”. It is composed of 3 words: “Epi” (meaning “upon”), “Demos” (meaning “people”), and “Logos” (meaning “the study of”), which, in all, means the study of what befalls a population [1]. For centuries, diseases were thought to appear due to superstition, myths, or religion and were often seen as a punishment [2]. In 400 B.C., Hippocrates (c. 460-c. 370 B.C.) was the first to see things more rationally and tried to demonstrate that diseases might occur and develop due to environmental and host factors [3]. It was not until the modern era and the middle of the 16th century that new hypotheses were introduced. Girolamo Fracastoro (1478-1553) was the first to suggest that diseases were caused by microscopic and unseeable particles or “spores”, able to spread by air or objects, to multiply by themselves, and be destroyed by fire [4,5]. Today, traditional epidemiology is the study of the distribution and determinants of health and disease in human populations, and applying of this knowledge to control health problems.

The roots of traditional epidemiology, as we know it today, can be traced back to studying infectious diseases in the 19th century. At the time, public health officials were concerned about the spread of diseases (such as cholera or tuberculosis), which were causing epidemics in different parts of the world. They began to collect data on these diseases’ incidence and prevalence and investigate their causes. A pioneer of modern epidemiology is John Snow (1813-1858), often called the father of epidemiology. In 1854, a cholera epidemic broke out in London, in the Soho district of London. Snow used observation and analysis of data to identify that the outbreak’s source was a contaminated water pump. His work demonstrated the importance of epidemiological methods in understanding and controlling infectious diseases [6,7].

As the major causes of illness and death were infectious diseases before the 20th century, epidemiology mainly focused on them. After John Snow's groundbreaking work identified the source of a cholera outbreak, public health conditions improved, and the overall burden of infectious diseases declined. Thus, epidemiology continued to evolve and expand as a new field of study on its own.

From acute diseases to chronic diseases

In the early 20th century, epidemiologists began investigating the causes of non-infectious diseases and their impact on public health. Non-infectious diseases (chronic or non-communicable diseases) are conditions characterized by long-term or slow progression that may get worse over time and are typically not caused by infectious agents [8]. Chronic diseases often have complex causes that involve genetic, environmental, and lifestyle factors, making them challenging to study [9,10]. Cancer, diabetes, heart disease, and chronic respiratory diseases are examples of chronic diseases. To investigate non-communicable diseases, epidemiologists tried using the same methods developed for infectious diseases, such as observational studies and clinical trials. For example, in the mid-20th century, epidemiological studies identified smoking as a major cause of lung cancer, leading to public health campaigns to reduce smoking rates [11,12].

Epidemiological methods

Another significant development of epidemiology was the evolution of epidemiological methods, such as case-control or cohort studies. Case-control and cohort studies are two types of observational studies used to investigate disease causes. In a case-control study, people with a disease (cases) are compared to people without the disease (controls) to identify factors that may have contributed to the disease [13]. A cohort study follows a group of people to see if certain factors are associated with developing a particular disease [14]. These methods were first developed in the mid-20th century and have since been refined and considered a gold standard in epidemiology [15].

Statistical Techniques and computing technologies

Statistical techniques play a critical role in epidemiology. They allow researchers to analyze data and conclude on disease’s causes and risk factors. Over the years, new statistical methods have been developed and refined to address the complex challenges of epidemiological research, such as confounding, bias, and small sample sizes [6].

The advent of computing technology in the mid-20th century has revolutionized epidemiology by allowing researchers to analyze and model large amounts of data [16,17]. Today, epidemiologists use various computational tools and software to manage, analyze, and visualize data, including statistical packages and programming languages like Python, R, or SAS, for multiple tasks such as modeling disease spread and progression [18].

Interdisciplinary

Epidemiology also became increasingly interdisciplinary. Epidemiologists started to collaborate with each other and health care professionals and providers, clinicians, and biostatisticians. Epidemiology is linked to health and biomedical, particularly biostatistics, but also to psychology, sociology, and economics [19,20]. The field also expanded to topics other than the study of infectious diseases or chronic diseases and included the investigation of health disparities or environmental health [21,22]. In particular, epidemiology plays a critical role in public health, helping to identify disease causes, track disease trends, and to develop interventions to prevent and control the disease. Epidemiologists are crucial players in the study of infectious diseases [23], non-infectious and chronic diseases such as cancer [24] and diabetes [25], and new emerging threats such as the COVID-19 pandemic [26]. With the increased use of the Internet on both computers and mobile phones, new online data are being generated that have enormous potential in epidemiology and for which new methods need to be developed or adapted.

Omics and digital epidemiology

Omics is a term that refers to some branches of science and technologies that aim to study and analyze biological systems on a large scale. It is derived from the suffix “-omics”, meaning a comprehensive study or analysis of a specific area. It involves using techniques to generate and analyze large datasets to gain insights into the structure, function and interactions of molecules in a system. Key omics disciplines are genomics, epigenomics, proteomics, metabolomics and microbiomics as seen in Figure 2.1. These often work together to provide a better understanding of complex biological systems. Omics have revolutionized biological research, personalized medicine and many other fields by enabling large-scale data generation, analysis and interpretation.

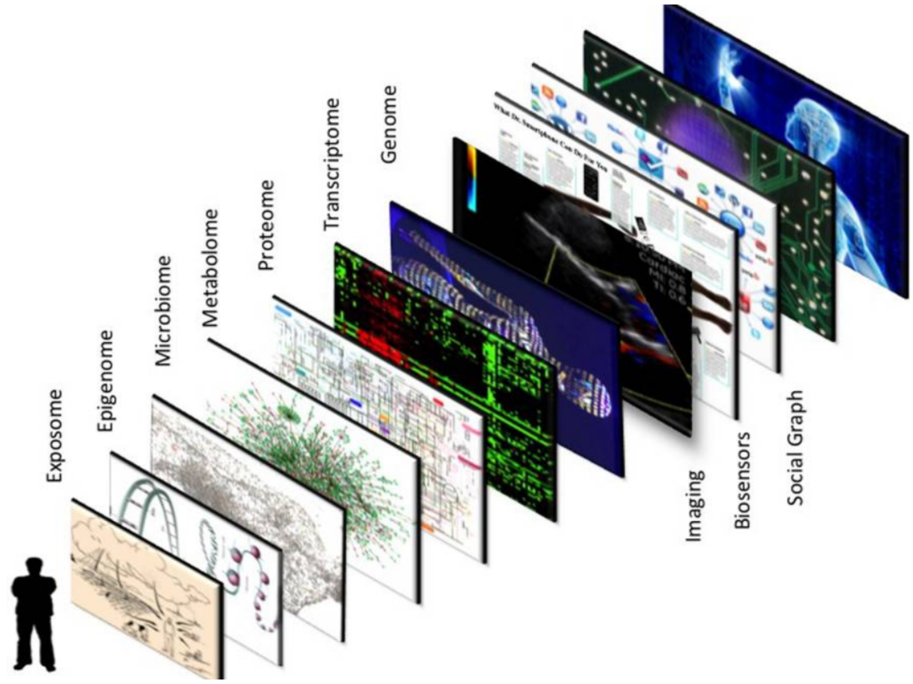


Figure 2.1. From Individualized Medicine from Prewomb to Tomb. Topol. Cell. 2014

A novel omic field has emerged during the last 30 years: the digitosome [27]. It encompasses integrating and analyzing data generated throughout an individual's life course from diverse digital sources such as wearables, smartphones, medical devices, and social media. These enable comprehensive insights into personal health profiles and behaviors. By leveraging data generated through digital sources, researchers can gain valuable insights into individual health profiles, disease patterns, risk factors, treatment outcomes, and trends at the global population level. Figure 2.2 shows how the digitosome can complement health research and impact prevention and patient care and management.

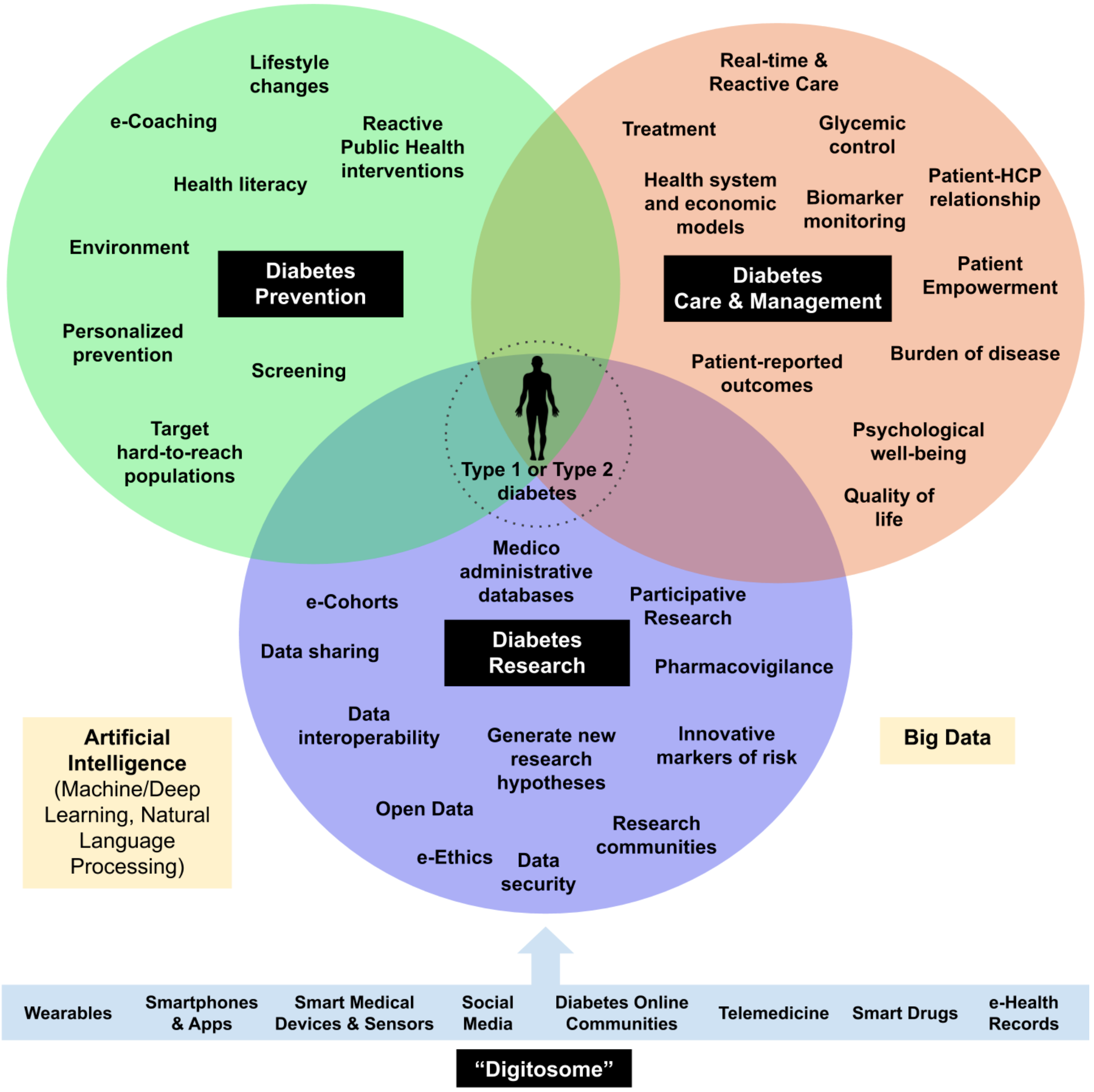


Figure 2.2. The digitosome: new technologies, data and artificial intelligence at the service of diabetes prevention, management, care and research. From Digital diabetes: perspectives for diabetes prevention, management and research. Fagherazzi et al. (Diabetes & Metabolism, 2018) - Open Access

In January 2023, nearly 65% of the world's population used the Internet [28] and more than two-thirds of the world’s total population used a mobile phone, 85% of them being smartphones [29]. Each request on the internet by a user leaves behind online data, more commonly known as a digital footprint. Digital footprints can be active or passive. An active digital footprint comprises all data shared by the users themselves (e.g. data from social media). In contrast, a passive footprint is created using all information about the users without them knowing about it (e.g. search engine queries) [30]. Such digital footprints have a high potential for health research. Digital epidemiology is thus defined as using digital data sources (social media data, internet search queries, mobile phone data..) that were not generated with the primary purpose of doing epidemiology to study and track the spread of diseases and public health trends [31]. Figure 2.3 shows the different recent and future medical innovations that can be used in digital epidemiology in the case of people with diabetes.

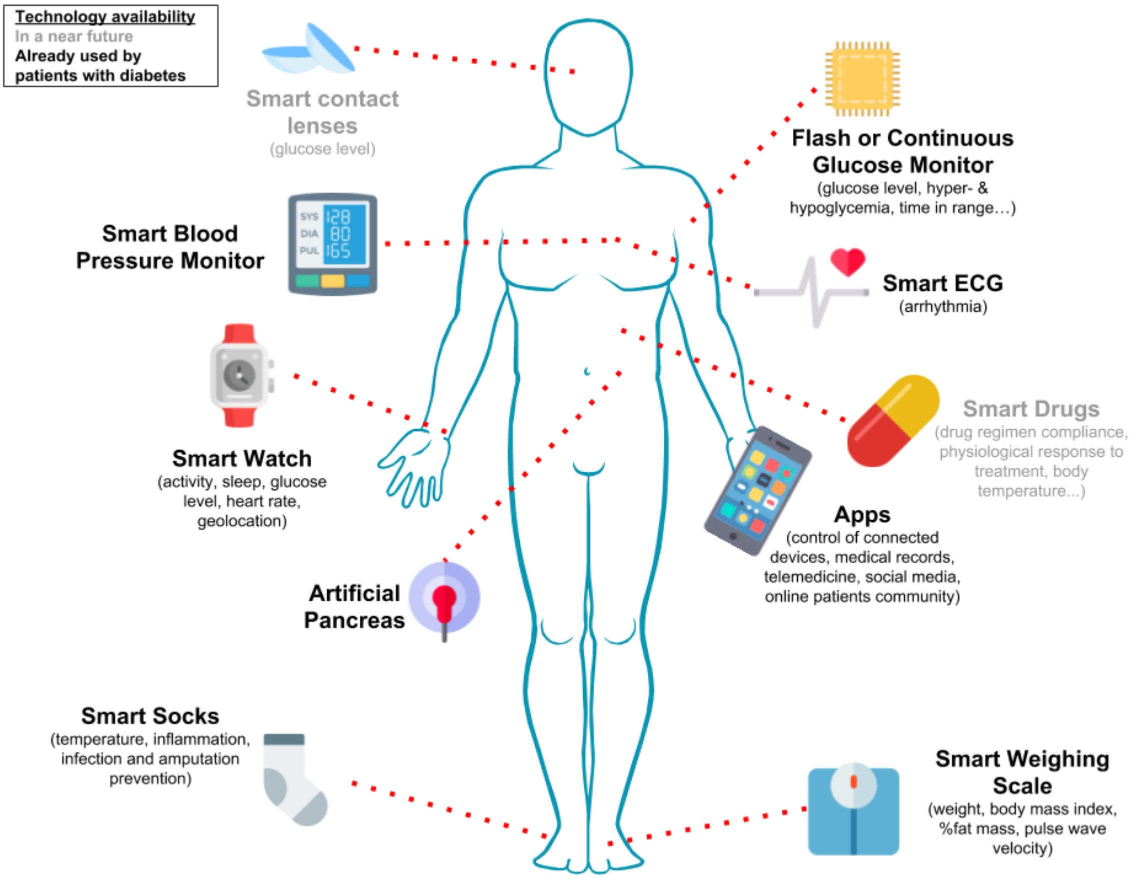

Technology availability
In a near future
Already used by patients with diabetes
Smart contact lenses
(glucose level)
Flash or Continuous Glucose Monitor
(glucose level, hyper- & hypoglycemia, time in range…)
Smart Blood Pressure Monitor
SYS 128
DIA 80
PUL 165
Smart ECG
(arrhythmia)
Smart Watch
(activity, sleep, glucose level, heart rate, geolocation)
Smart Drugs
(drug regimen compliance, physiological response to treatment, body temperature...)
Apps
(control of connected devices, medical records, telemedicine, social media, online patients community)
Artificial Pancreas
Smart Socks
(temperature, inflammation, infection and amputation prevention)
Smart Weighing Scale
(weight, body mass index, %fat mass, pulse wave velocity)

Figure 2.3. Recent and future medical innovations to help people living with diabetes. ECG: electrocardiography. From Digital diabetes: perspectives for diabetes prevention, management and research. Fagherazzi et al. (Diabetes & Metabolism, 2018) - Open Access

The beginning of digital epidemiology dates back to the early 2000s when researchers used search engine query data from Google to track infectious influenza epidemics [32]. However, the field began to gain traction in the mid-2010s with the increasing availability of big data and the development of machine learning and data analytics techniques.

However, it is important to note that even if traditional and digital epidemiology both aim to understand and identify patterns and determinants of health, they also have several differences in their methodologies and limitations. Traditional epidemiological studies often use a specific selection process to make sure the participants will be a representative sample of the studied population [33,34]. This will allow more controlled analyses and offer results that can be generalized to the population from which the sample was drawn. In contrast, digital epidemiology, while offering the opportunity to collect analyze large amounts of data in real-time, struggles to ensure its data sources accurately represent a whole population [35]. Social media data, for instance, may exclude certain demographic groups, such as the elderly or those without internet access, potentially leading to biased perspectives. Furthermore, the passive collection of digital footprints can introduce biases, as individuals might not always be honest online or might leave out important details. Still, such data can also be considered as less subject to biases than reports collected in a traditional setting [36]. Moreover, the variability and dynamic nature of online platforms and digital tools can sometimes make standardization and comparability challenging. Digital epidemiology is fast and covers a wide range of data, but ensuring it represents everyone is tough. Thus, by combining it with traditional methods, we can get a a more comprehensive picture of health issues and associated research questions, using the strengths of both and addressing their weaknesses.

In this thesis, we try to mirror the transition that happened in traditional epidemiology using online data and AI to study chronic diseases such as diabetes, all while considering and addressing the aforementioned limitations in our analysis, results and conclusions.

The need for more minimally disruptive health research

We chose the research framework of minimally disruptive health research to guide us through this project. In healthcare, individuals may face up to three types of burdens: the burden of disease, burden of treatment and burden of research.

Minimally disruptive medicine is a concept that aims to reduce such burdens on patients' lives by minimizing the disturbance caused by medical treatments and by tailoring the treatments to patients' experiences, preferences and goals [37,38]. Figure 2.4 shows this balance between treatment and disease burdens and how it can easily be unbalanced. While disease burden is the impact of an illness on a patient's daily life, treatment burden refers to the mental and physical effort required to manage the treatment. Still, both these burdens have an impact on the quality of life. However, the research burden is another challenge when dealing with a chronic condition, including the inconveniences that patients might experience when participating in clinical research. In this thesis, the hypothesis was that, thanks to social media data, we can contribute to the reduction of the research burden by providing real-time insight into patient experiences and concerns, reducing the need to use questionnaires and facilitating patient-centric research.

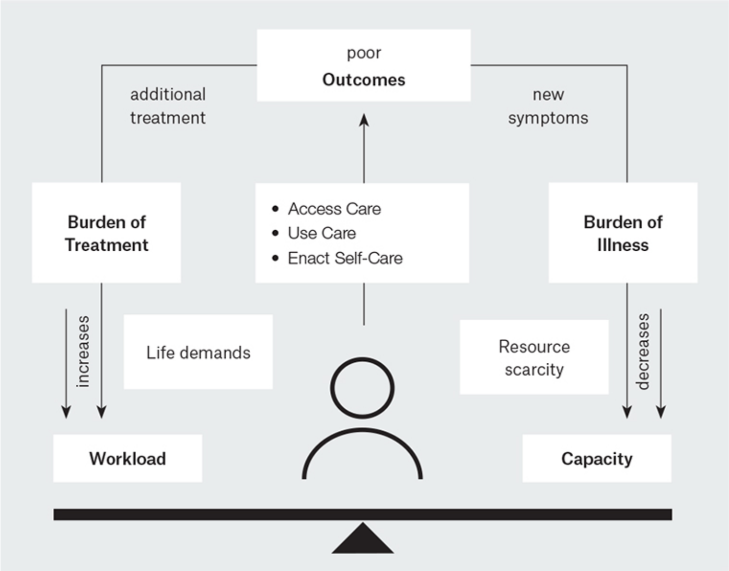
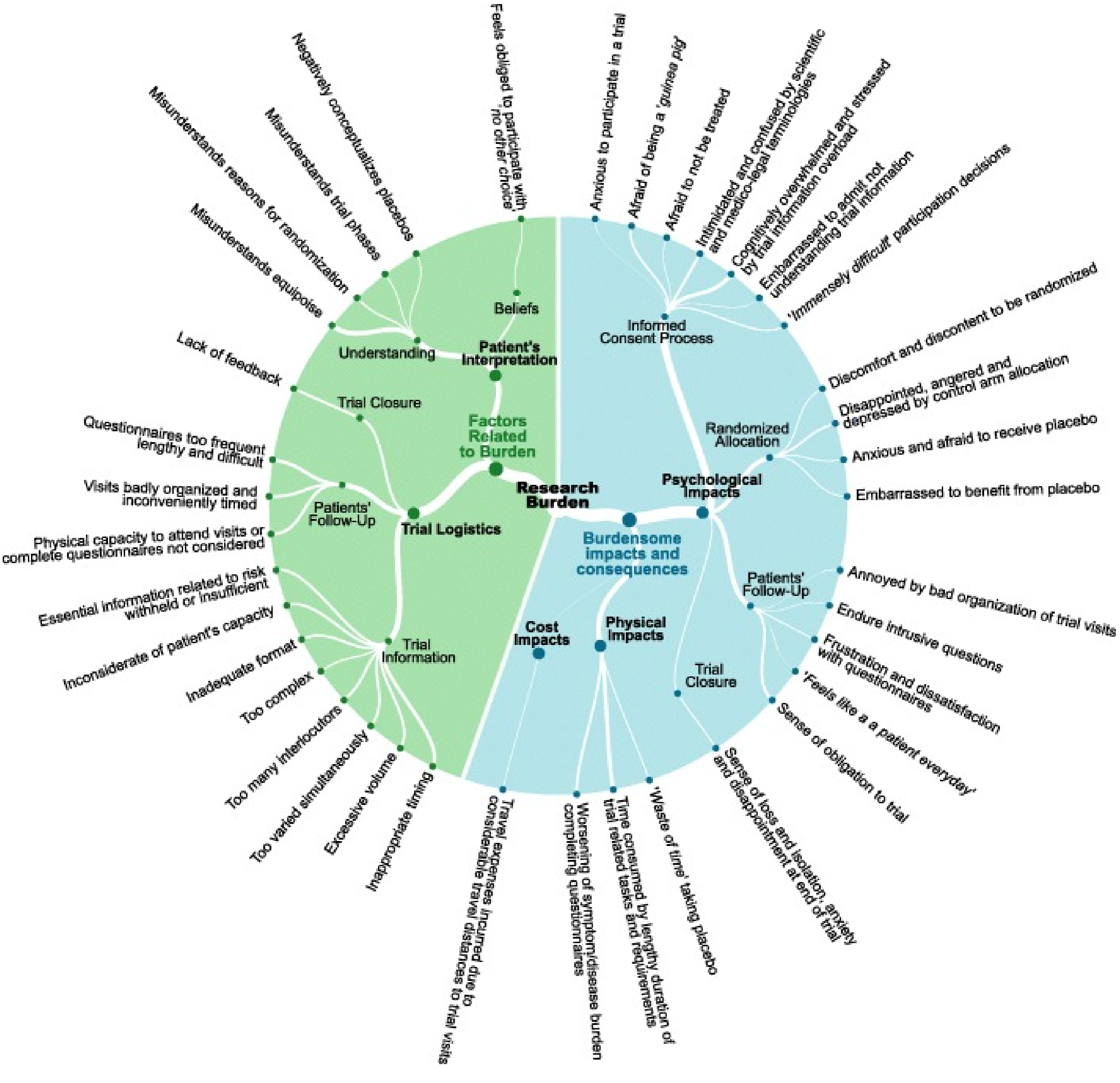


Figure 2.4. The cumulative complexity model. Serrano, V., Spencer-Bonilla, G., Boehmer, K.R. et al. Minimally Disruptive Medicine for Patients with Diabetes. Curr Diab Rep 17, 104 (2017).

Healthcare providers and researchers aim to find a balance that optimizes patient outcomes while minimizing the overall burden of treatment and disease. This can involve a patient-centered approach that considers the patient's preferences, goals, and the impact of their illness and treatment burden on their daily life. Thus, by finding the right balance, patients can better manage their illnesses, improve their quality of life, and achieve better health outcomes.

Besides these two types of burden, we usually add a third burden for patients who take part in health research. Indeed, patients' participation in clinical or epidemiological research can counteract this balance and generate an additional "research burden".

Patients may need to travel, visit healthcare professionals more often, undergo supplementary procedures or complete questionnaires that will be time-consuming [39]. Figure 2.5 highlights all the challenges linked with the research burden, showing the need for minimally disruptive research.



Research Burden
Factors Related to Burden
Burdensome impacts and consequences
Patient's Interpretation
Beliefs
Understanding
Trial Logistics
Trial Closure
Patients' Follow-Up
Trial Information
Informed Consent Process
Psychological Impacts
Randomized Allocation
Patients' Follow-Up
Trial Closure
Cost Impacts
Physical Impacts
Feels obliged to participate with 'no other choice'
Negatively conceptualizes placebos
Misunderstands trial phases
Misunderstands reasons for randomization
Misunderstands equipoise
Lack of feedback
Questionnaires too frequent lengthy and difficult
Visits badly organized and inconveniently timed
Physical capacity to attend visits or complete questionnaires not considered
Essential information related to risk withheld or insufficient
Inconsiderate of patient's capacity
Inadequate format
Too complex
Too many interlocutors
Too varied simultaneously
Excessive volume
Inappropriate timing
Anxious to participate in a trial
Afraid of being a 'guinea pig'
Afraid to not be treated
Intimidated and confused by scientific and medico-legal terminologies
Cognitively overwhelmed and stressed by trial information overload
Embarrassed to admit not understanding trial information
'Immensely difficult' participation decisions
Discomfort and discontent to be randomized
Disappointed, angered and depressed by control arm allocation
Anxious and afraid to receive placebo
Embarrassed to benefit from placebo
Annoyed by bad organization of trial visits
Endure intrusive questions
Frustration and dissatisfaction with questionnaires
'Feels like a a patient everyday'
Sense of obligation to trial
Sense of loss and isolation, anxiety and disappointment at end of trial
'Waste of time' taking placebo
Time consumed by lengthy duration of trial related tasks and requirements
Worsening of symptom/disease burden completing questionnaires
Travel expenses incurred due to considerable travel distances to trial visits

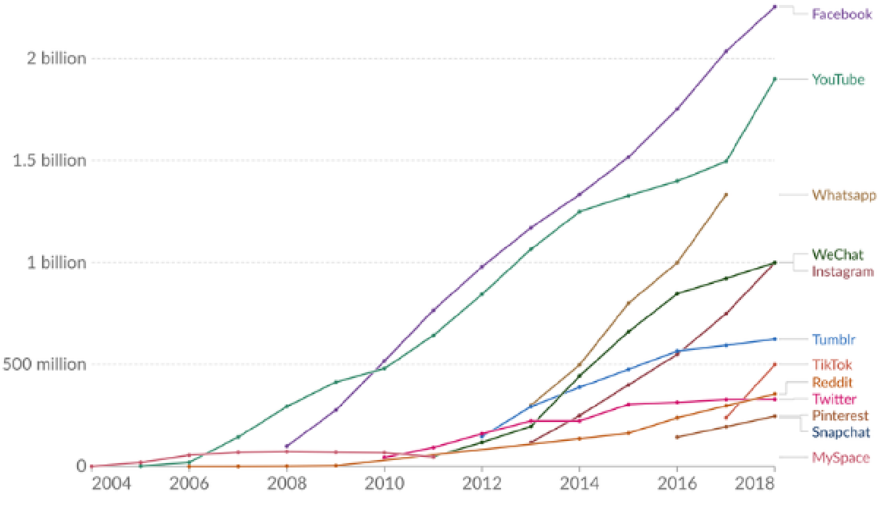
Figure 2.5. Classification of theme "Research Burden" and sub-themes "Burdensome Impacts and Consequences" and "Factors related to Burden". From Naidoo N, Nguyen VT, Ravaud P, Young B, Amiel P, Schanté D, Clarke M, Boutron I. The research burden of randomized controlled trial participation: a systematic thematic synthesis of qualitative evidence. BMC Med. 2020 Jan

Social media and health research

Social media

Social media were created in the early 2000s [40] when online forums and chat rooms emerged. Social media are internet and mobile phone-based tools that enable users to create, share, and exchange text, images, videos, music, and information with each other [41]. One of the earliest and most influential social media platforms was Friendster, launched in 2002. Other early platforms included MySpace (2003), LinkedIn (2002), and Facebook, initially launched in 2004 as a platform for college students.

Since then, social media have exploded in popularity, with the number of platforms and users growing exponentially. Twitter (now X), Instagram, Snapchat, TikTok, and, more recently Threads are just a few of the many social media that have emerged in recent years, each with its own features and user bases. Figure 2.6 shows the evolution of the number of social media users from 2004 to 2018.



Source: Statista and TNW (2019)

OurWorldInData.org/internet • CC BY

Figure 2.6. Number of people using social media from 2004 to 2018 based on monthly active users. Taken from OurWorldInData.org

Today, social media are an integral part of modern communication and culture, with billions of people worldwide using them daily to connect with others, share their experiences, or consume content.

Social media and health research

As mentioned earlier, internet users leave a digital footprint with immense potential for health research. Indeed, social media are a valuable source of health-related data. In 2010, 11% of social media users in the United States posted comments, queries, or information about health or medical content [42]. With all this health-related data being generated and relatively easily accessible, health researchers have started to take an interest and use them for various tasks such as recruitment and data collection [43].

PubMed is a search engine on the MEDLINE database of references and abstracts on life sciences and biomedical topics. On PubMed, Medical Subject Headings (MeSH terms) are a vocabulary used to index journal articles and books. “Social Media” became a MeSH term in 2012, showing the growing interest in social media for biomedical and life sciences research. Figure 2.7 also highlights the number of publications per year using the MeSH term “Social Media” since its introduction in 2012.

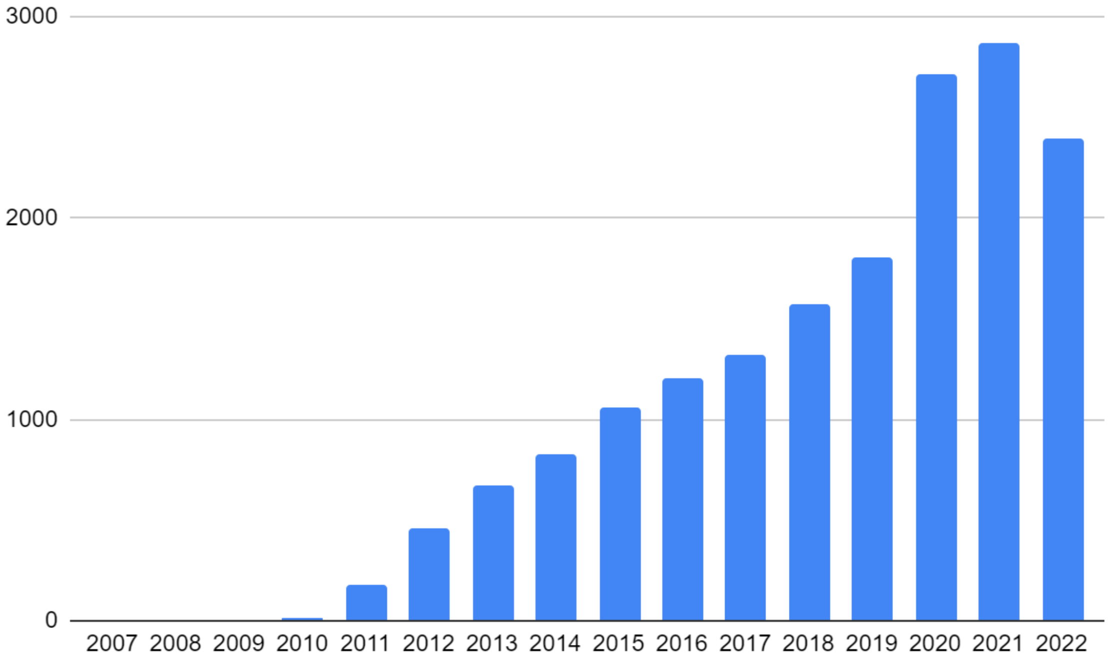


Figure 2.7. Evolution of the number of articles and references with the MeSH term “Social Media” between 2007 and 2022.

It can be noted that during the COVID-19 pandemic in 2020 and 2021, social media were mainly used for health research, which allowed for the rapid development of new analysis methods.

The case of Twitter

Twitter (now called X) is a popular social media platform created in 2006 that enables users to share short messages, called tweets, with their followers. With over 330 million active users worldwide, Twitter has become a powerful tool for communication, social interaction and news dissemination. On average, more than 500 million tweets are published each day. Tweets can contain text, images, and links. Users can engage with each other by liking, retweeting, and replying to tweets. Between 2006 and early 2023, Twitter emerged as a valuable tool for health research, mostly due to its freely available Application Programming Interface (API). This API, detailed further in Chapter 3 (Twitter API), facilitated the collection and access of data for researchers. However, in 2023, Twitter evolved into 'X', and the corresponding X API was no longer freely accessible. It is important to note that all the research and methodologies presented in this thesis were developed in the pre-2023 Twitter API and the platform known as Twitter. Still, the principles and findings remain applicable to 'X', as long as one is willing to invest in accessing its API. Thus in this thesis, we will continue to use and reference Twitter as our primary platform for analysis. Besides, we address later in the thesis how the methodology and technologies developed in this work can be transferred to other social media platforms and data sources.

On social media, particularly on Twitter, engaged online communities of patients have emerged. In particular, two of the most active supportive networks for individuals living with diabetes were created with, for instance, the GBDoc (Great Britain Diabetes online community with 5,500 followers on @theGBDOC) and DSMA (Diabetes Social Media Advocacy with 20,200 followers on @DiabetesSocMed) groups. These communities are valuable for sharing experiences, seeking guidance and fostering connections among people facing similar challenges related to diabetes management. Hashtags such as #GBDoc, #DSMA, #DiabetesCommunity, and #T1D (Type 1 Diabetes) serve as digital markers, enabling users to find and participate in conversations on topics like blood glucose monitoring, insulin dosing, dietary considerations, exercise routines, emotional well-being, and the latest advancements in diabetes research and treatment. The power of social media and online diabetes communities lies in their ability to overcome geographic boundaries, enabling individuals with different backgrounds and locations to connect and exchange knowledge. These digital spaces have fostered a sense of empowerment and solidarity within the community of people with diabetes, amplifying voices and shaping the narrative around diabetes care and management.

Diabetes

Below we define crucial diabetes-related concepts that are then analyzed thanks to social media in this thesis.

Diabetes is a chronic metabolic disorder characterized by high blood sugar levels (also called hyperglycemia) resulting from the body's inability to produce or properly use insulin [44]. The impact of diabetes is significant, with a growing number of cases worldwide and being ranked as the 9th cause of death worldwide by the WHO [45]. Currently, there are an estimated 536.6 million adults aged 20-79 with diabetes, and this number is expected to continue rising, as shown in Figure 2.8. The symptoms of diabetes can include frequent urination, excessive thirst, hunger, fatigue, blurred vision, and slow healing of wounds. These can lead to severe complications, such as heart disease, kidney disease, nerve damage, and retinopathy [46].

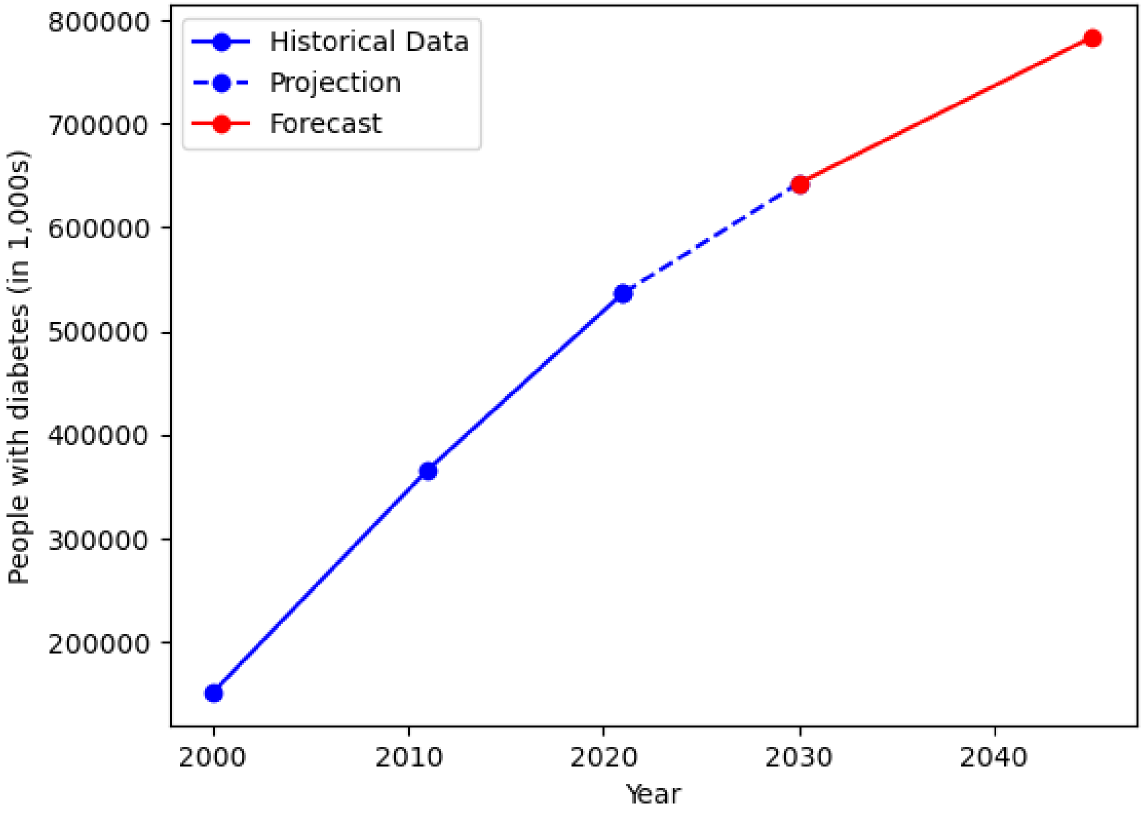


Figure 2.8. Evolution and forecasting of the number of adults with diabetes from 2000 and 2045. Data from https://diabetesatlas.org/data/en/world/

Understanding the differences between Type 1 and Type 2 diabetes, the two main types of diabetes, is critical for a proper diagnosis, treatment, and management. While both types of diabetes impact blood sugar levels, they differ in their causes, age of onset, and treatment approaches. Regular monitoring, adherence to prescribed treatments, and ongoing support are essential for individuals with either type of diabetes to live healthy lives.

Type 1 diabetes

Type 1 diabetes is a chronic autoimmune disorder characterized by the destruction of pancreatic beta cells, leading to a complete deficiency of insulin secretion [47]. This condition is typically diagnosed during childhood or adolescence but can also develop in adults. Type 1 diabetes affects approximately 5-10% of all diagnosed cases of diabetes globally [48]. The exact cause of type 1 diabetes is not fully understood yet, but it is thought to be a combination of environmental and genetic factors [49]. The incidence of type 1 diabetes is increasing, and the prevalence is expected to blow in the coming years [50,51].

The symptoms of type 1 diabetes can develop rapidly over a few days to weeks and include increased thirst and urination, fatigue, weight loss, and blurred vision. In some cases, diabetic ketoacidosis can occur, which is a life-threatening condition characterized by the accumulation of ketones in the blood, leading to metabolic acidosis [52].

The diagnosis of type 1 diabetes is based on clinical symptoms and laboratory tests. Fasting plasma glucose, oral glucose tolerance tests, and glycosylated hemoglobin (HbA1c) levels are commonly used to diagnose diabetes [53–55]. In addition, tests for autoantibodies such as islet cell antibodies, glutamic acid decarboxylase antibodies, and insulin autoantibodies can help to confirm the diagnosis of type 1 diabetes [56].

The management of type 1 diabetes is focused on insulin therapy, which aims to replace insulin secretion deficiency [57]. Insulin can be administered via injection (pen or syringe) or insulin pump [58]. In addition to insulin therapy, lifestyle modifications such as healthy eating and regular exercise are essential in managing type 1 diabetes [59]. Continuous glucose monitoring and self-monitoring of blood glucose can help to optimize insulin therapy and prevent hypoglycemia [60].

Type 2 diabetes

Type 2 diabetes is a chronic disorder characterized by high blood glucose levels due to insulin resistance and impaired insulin secretion by the pancreatic beta cells [61]. Type 2 diabetes is the most common type of diabetes and is responsible for approximately 90%-95% of all diagnosed cases of diabetes [62]. The prevalence of type 2 diabetes has been increasing rapidly and is expected to rise dramatically in the future [63]. It is a complex disease that results from the interplay of genetic, environmental, and lifestyle factors. Common environmental factors are a sedentary lifestyle, an unhealthy diet, and obesity, which contribute to developing type 2 diabetes [64].

The symptoms of type 2 diabetes include increased thirst and frequent urination, fatigue, blurred vision, slow wound healing, numbness or tingling in the hands or feet, recurrent infections, and unexplained weight loss [65]. These symptoms may vary in severity and onset, and some people with type 2 diabetes may not experience any symptoms, particularly in the early stages of the disease.

The diagnosis of type 2 diabetes is based on measuring fasting plasma glucose or HbA1c levels. Fasting plasma glucose levels greater than or equal to 126 mg/dL (7.0 mmol/L) or HbA1c levels greater than or equal to 6.5% indicate the presence of diabetes [66].

The management of type 2 diabetes involves a combination of lifestyle modifications, pharmacotherapy, and monitoring of blood glucose levels. Lifestyle modifications such as regular physical activity, a healthy diet, and weight loss can improve insulin sensitivity and the glucose uptake [67].

Other types

While type 1 and type 2 diabetes are the most well-known forms of the disease, there are several less common types of diabetes.

Gestational diabetes is a type of diabetes that affects pregnant women. It is a type of diabetes that occurs during pregnancy and affects about 2-10% of pregnant women [68]. This is caused by hormonal changes that affect the body's ability to produce and use insulin effectively. The symptoms of gestational diabetes are similar to those of type 2 diabetes and may include increased thirst and urination, blurred vision, fatigue, and frequent infections. However, many women with gestational diabetes may not experience any symptoms, which is why screening is important. It usually resolves after giving birth but in some cases it can turn to type 2 diabetes [68]. Gestational diabetes is typically diagnosed between the 24th and 28th week of pregnancy through a glucose tolerance test. This involves drinking a sugary drink and then having blood drawn to measure blood glucose levels [69]. Treatment for gestational diabetes usually involves dietary changes, regular physical activity, and monitoring blood glucose levels. In some cases, medication or insulin therapy may also be required to manage blood sugar levels [70].

Finally, there are also rarer types of diabetes such as monogenic diabetes [71] or the latent autoimmune diabetes in adults [72].

Diabetes management

Blood glucose monitoring

People with diabetes need to monitor their blood sugar levels regularly. There are different ways to measure blood sugar levels; the most common is using a blood glucose meter. The patient needs to prick their finger with a lancet to obtain a drop of blood. Then, they place the blood on a test strip and insert it into the meter. The meter measures the blood glucose amount and displays the result on its screen. Based on the results the person can decide their insulin dosage, food intake, and other aspects of diabetes management. It is also possible to use continuous glucose monitoring systems. It involves wearing a small device that measures blood sugar levels. The device sends the information to a smartphone or another device, allowing the person to monitor their glucose levels in real-time throughout the day. Continuous glucose monitoring systems can be beneficial for people with type 1 diabetes, as they can prevent hypoglycemia and hyperglycemia [73].

Insulin

As mentioned earlier, insulin is used to treat diabetes. Insulin deficiency or resistance can lead to diabetes [74]. Insulin deficiency is a condition where the body doesn’t produce enough insulin or doesn’t use effectively the insulin it produces.

Insulin is a hormone usually produced by the pancreas, an organ located behind the stomach that plays a crucial role in regulating the body's blood glucose levels. Insulin is released in response to increased blood glucose levels after meals [75]. Then, it travels through the body and attaches to specific receptors in the muscles. This triggers a reaction that allows glucose to enter the cells. Glucose will be directly used for energy or stored for future use [76]. Insulin regulates the production of glucose and the breakdown of glucose in the liver. When blood glucose levels are high (hyperglycemia), insulin is going to indicate to the liver to store excess glucose as glycogen. When blood glucose levels are low (hypoglycemia), insulin signals the liver to break down the glycogen and release glucose into the bloodstream [77].

Thus, in the case of diabetes, insulin needs to be administered. Such insulins can be short-acting, rapid-acting, intermediate-acting, long-acting or mixed. There are several ways to administer insulin, such as subcutaneous injections, insulin pumps, and inhalation therapies. Subcutaneous injection is the most common one and involves injecting insulin into the layer of fat just below the skin. Insulin can be injected using a syringe or an insulin pen, which is a device that contains a cartridge of insulin and a needle [78]. Insulin pumps involve wearing a small, portable pump that is going to automatically and continuously deliver insulin throughout the day. The pump has to be manually filled with insulin. Then, it will release small amounts of insulin into the body at regular intervals. The user can also manually administer extra doses of insulin when needed [79]. Inhalation therapy involves using an inhaler to deliver insulin into the lungs, which is then absorbed into the bloodstream. This method is less common than subcutaneous injection or insulin pump therapy, but it may be a suitable option for some people with difficulty with injections [80]. Ultimately, the method of insulin administration will depend on the individual's needs and preferences, as well as their healthcare provider's recommendations.

Diabetes burden

Diabetes burden is described by all the physical, emotional, interpersonal, regimen-related and financial challenges of managing and living with diabetes, as shown in Figure 2.9. It includes, for instance, the need to monitor blood glucose levels, take medication, adhere to a strict diet and exercise regimen, which are daily concerns and make decisions related to those [81]. It can also include the impact of the disease on an individual's social life and relationships, as well as its economic burden due to the cost of healthcare and related expenses [82]. All these create a mental load on people with diabetes that patients try to balance using coping mechanisms, external support (e.g. family and friends) or with the help of healthcare professionals.

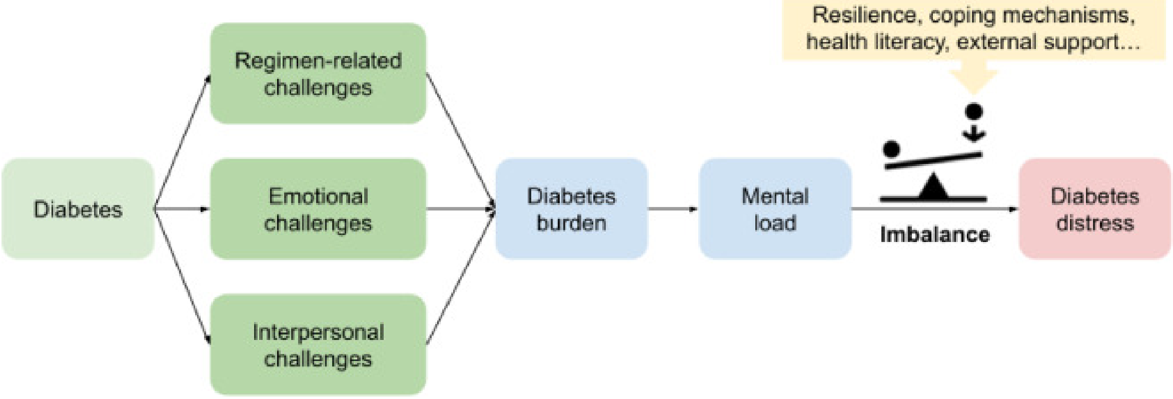


Figure 2.9. Simplified conceptual framework between diabetes and diabetes distress. From Guy Fagherazzi, Technologies will not make diabetes disappear: how to integrate the concept of diabetes distress into care, Diabetes Epidemiology and Management, Volume 11, 2023, 100140, ISSN 2666-9706,

Diabetes distress

Diabetes distress refers to the emotional and psychological impact of living with diabetes. It includes feelings related to the daily management of diabetes, such as anxiety, depression, and frustration [83]. It can also include worries about the potential complications of diabetes and the impact of the disease on the quality of life [84]. Diabetes distress can significantly impact patients, leading to decreased quality of life but also increased risk for complications [85]. People with diabetes distress can experience a wide range of symptoms such as sleep disturbances, fatigue, changes in appetite, difficulty concentrating, less motivation to manage their disease, etc [86]. These can lead to treatment nonadherence (e.g. missed insulin dose), which can worsen blood glucose control and increase the risk of long-term complications [87].

Diabetes distress is commonly evaluated using self-reported questionnaires, with the most used ones being the Diabetes Distress Scale (DDS) and the Problem Areas in Diabetes (PAID) questionnaire. The PAID questionnaire consists of 20 items that explore negative emotions commonly experienced by individuals with diabetes, such as fear, anger, and frustration. More generally, these questionnaires aim to assess the emotional and psychological burden experienced by people with diabetes. However, these questionnaires have certain limitations. First, they are non-evolutive measures, meaning they capture diabetes distress during the questionnaire and may not reflect changes over time. Second, some components of diabetes distress are missing, such as work-related issues, cost of treatment and healthcare-providers relationships. Third, patients may provide answers they believe healthcare providers want to hear rather than expressing their true feelings, also called the “make my doctor happy” effect. Fourth, these questionnaires rely on self-reporting, which can be influenced by subjective interpretation and recall bias. Finally, patient’s perception of their own distress levels may differ from objective assessments, potentially leading to an incomplete understanding of their emotional state.

Reaching an effective management of diabetes distress requires a multifaceted approach that addresses physical and emotional aspects. Treatment may include diabetes education, psychological counseling, and support groups. Another critical factor is to reduce the diabetes burden by using technology to develop self-care strategies that prioritize patient preferences and values. There is also a major need to describe further and increase awareness around diabetes distress among healthcare providers to improve the detection of mental health issues or difficulties related to diabetes management in consultations.

By addressing diabetes distress and improving patient well-being, healthcare providers can help patients to achieve better diabetes outcomes and improve the overall quality of life of people with diabetes.

Social media is a rich data source for studying online diabetes communities and an opportunity to gain insights into patient experiences, identify unmet needs, and inform the development of targeted interventions and support services. By analyzing the discussions, sentiments, and information shared within these communities, researchers can uncover trends, identify emerging issues, and contribute to the knowledge surrounding diabetes management and patient well-being. Ahne et al. showed that social media data can be a valuable source of information to identify the key concerns of people with diabetes and thus better address these concerns [88]. This thesis aims to further investigate the understanding of diabetes-related distress from the perspective of people with diabetes through their social media data.

Chapter 3

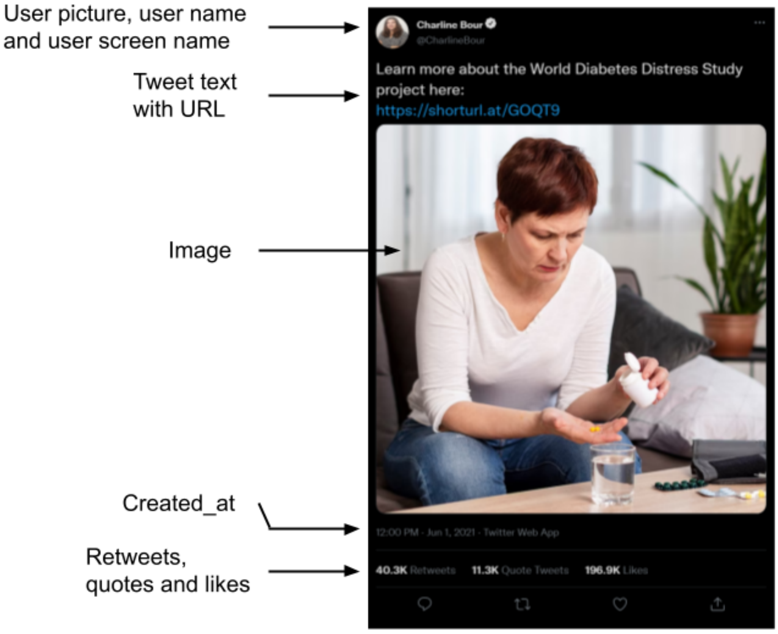
Material and methods

This chapter describes the principal materials and methods used in this study to analyze Twitter data using Natural Language Processing (NLP) techniques and artificial intelligence.

While each objective requires specific methods and tools, some techniques are reused across objectives, such as data collection and preprocessing, which will not be described in detail each time. Instead, we will focus on describing the relevant materials and methods in a structured and concise manner. We provide a detailed description of each objective, including the data sources and collection, the data preprocessing and cleaning, the modeling, statistical analysis, and visualization.

Twitter

Twitter is a social media that allows users to share short messages called tweets. It is a real-time public conversation platform where individuals, organizations, and public figures can communicate with a broad audience. Initially, tweets were limited to 140 characters, but it changed in 2017 to 280 characters. Recently, Twitter introduced Twitter Blue, a paid subscription providing new features such as writing tweets up to 4,000 characters long. No matter how long it is, a tweet can contain text, images, videos, links, and hashtags. Hashtags are words or phrases preceded by the '#' symbol. They are clickable links, allowing users to discover and follow discussions related to a specific topic. Hashtags help create a sense of community and facilitate the navigation of relevant content across social media platforms. Each user has a chronological feed of posts or updates called a timeline. Users can follow other accounts to see their tweets in their timelines and engage with tweets by liking, retweeting, or replying to them. Figure 3.1 is an example of a tweet.


User picture, user name and user screen name
Tweet text with URL
Image
Created_at
Retweets, quotes and likes

Charline Bour
@CharlineBour

Learn more about the World Diabetes Distress Study project here:
https://shorturl.at/GOQT9

12:00 PM · Jun 1, 2021 · Twitter Web App

40.3K Retweets 11.3K Quote Tweets 196.9K Likes

Figure 3.1. Example of a factice tweet and its metadata. (generated with Tweetgen)

Twitter API

The Twitter API (Application Programming Interface) is a set of rules and protocols that allows developers to interact with Twitter's platform and access its data [89]. It provides a way to retrieve, post, and interact with tweets, user profiles, trends, and more. The API offers several endpoints corresponding to different functionalities, such as posting new tweets, retrieving user information, searching for tweets, and accessing user timelines. By integrating the Twitter API into their applications, developers can create custom solutions, analyze data, build social media management tools, and enable interactions with the Twitter platform. There are two versions of this API. The first one released was v.1.1 (used in objective 2) which recently became deprecated and the v.2 (which was used in objective 3).

Access to the API depends on creating a Twitter Developer Account through the Twitter Developer Portal. A valid Twitter account, a project name, and a description of how the API and data will be used are required. Developers are also asked to provide a use case category, such as business, public, or research, based on the purpose of their next application. It should be noted that different use case categories have varying environments (e.g., free, basic, enterprise, or academic), each offering other rights and freedoms regarding API usage. The application is then submitted, subject to review, and it may take some time before receiving feedback from Twitter. Once approved, the Twitter Developer Portal can be accessed. A Twitter App can be created for a project, specifying the use case. For data collection, the permission should be “Read-Only”. The API keys and access tokens can be generated, which will need to be integrated into the Python scripts to stream and collect the data.

However, the pricing and access rights for academic use cases may change. It is essential to stay updated with the latest information and guidelines provided by Twitter regarding API access and associated costs.

Data collection

In Python, the Tweepy library can stream Twitter and collect tweets [90]. The filter stream endpoint of Tweepy was used in objective 2 to collect diabetes-related tweets based on a list of keywords. Tweets that matched specific keywords related to diabetes, such as "diabetes", "insulin," or "hypoglycemia" were retrieved. We also expanded the list of keywords to include terms in different languages to filter tweets from various countries and try to cover a large part of the world. These keywords can be seen in Appendix 3.

Metadata

When collecting tweets with the stream filter, various metadata is accessible. This includes information about the tweets and the users, such as the tweet ID, text content, creation time, source application, language, associated hashtags, mentions, URLs, media attachments, author information (username, user ID, screen name, bio), user location, user followers count, verified status, and profile image URL. Additionally, engagement-related metadata is available, such as retweets and favorite counts. Finally, if enabled, geolocation metadata can provide information about the precise coordinates or place associated with a tweet, but tweets with this metadata are more rare.

Data storage

MongoDB is a NoSQL database system used to store data [91]. It is a document-oriented database where data is stored as JSON-like documents called BSON. These documents are gathered into collections. With its rich features and ability to handle large volumes of data, MongoDB is used in modern applications for efficient and scalable data storage and retrieval. Figure 3.2 shows the data flow between the API and MongoDB.

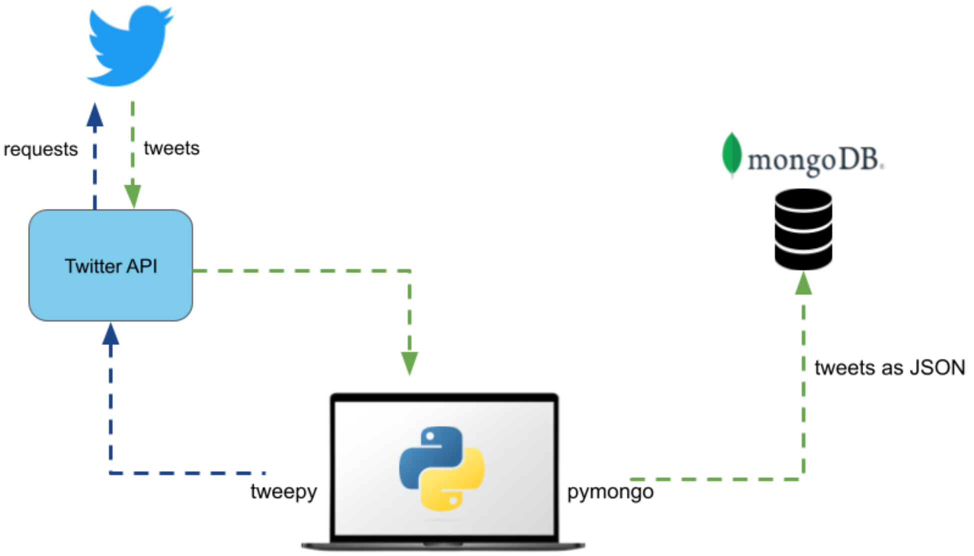


Figure 3.2. Data Flow Process: Retrieving Twitter Data and Storing in MongoDB

Pandas

Pandas is an open-source library in Python for data manipulation and analysis [92]. It allows the user to easily load, manipulate, and analyze structured data, perform data cleaning and transformation tasks, handle missing values, apply filtering and grouping operations, and perform statistical computations. Pandas is a tool for data scientists that significantly simplifies the data processing workflow and enables insightful data exploration and analysis. It was used to manipulate the data for most of the preprocessing steps. This library was mostly used to manipulate and preprocess the tweets.

Natural Language Processing

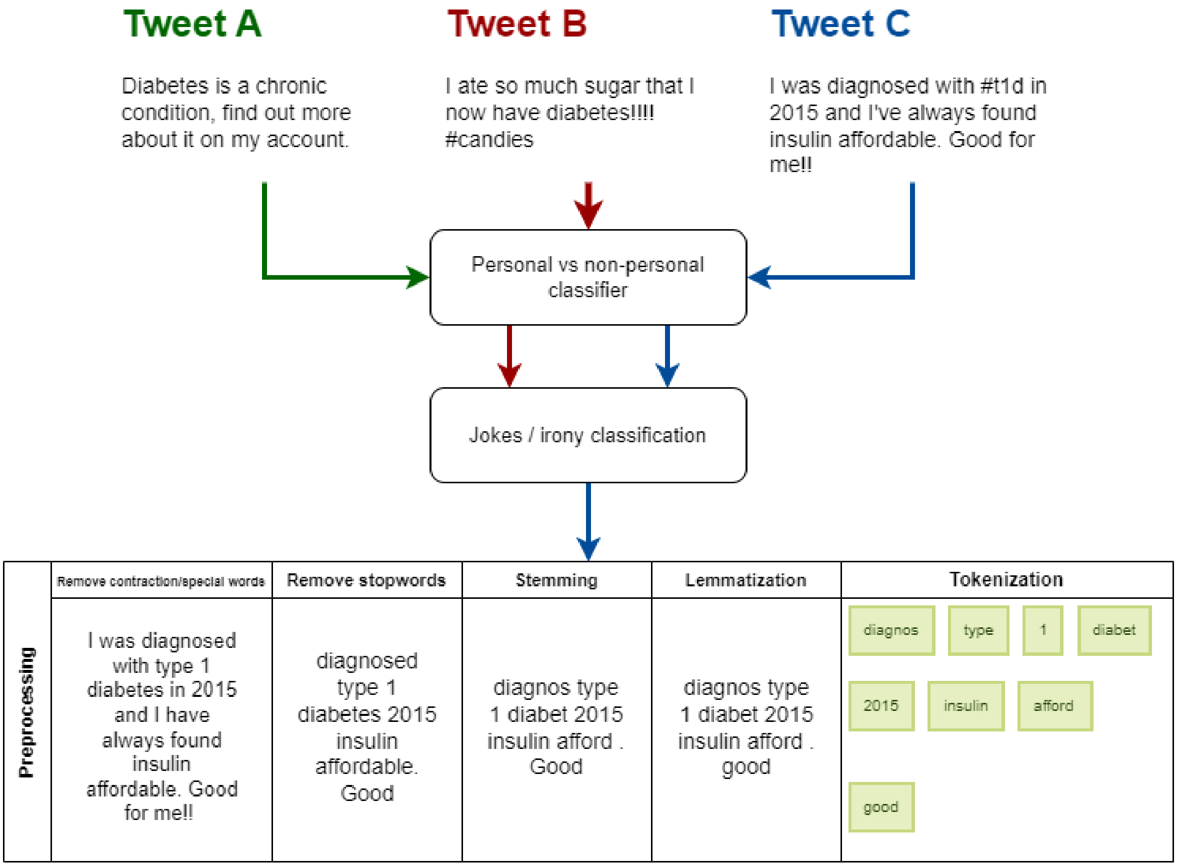
Natural Language Processing (NLP) is a branch of artificial intelligence that involves the development of algorithms that allow computers to understand and interpret in a way that is both meaningful and contextually relevant. NLP includes various tasks, including text classification, sentiment analysis, translation, named entity recognition, question answering, and text generation. It plays a vital role in applications such as chatbots, virtual assistants, language translation, information retrieval, and text mining, enhancing human-computer interaction and enabling machines to comprehend and communicate in natural language.

Translation

Translating all tweets to the same language, here English, is an important step. It allows standardized and consistent input for the algorithms, enabling them to learn patterns and make accurate predictions effectively. While translating may result in the loss of specific linguistic nuances, it enables a more comprehensive data analysis, especially when dealing with large datasets or when leveraging pre-trained English-specific models. The translation was performed in objective 2 using Google Translate API on the package deep-translator [93].

Preprocessing steps

In NLP, several typical preprocessing steps are employed to transform raw text into a more suitable format for analysis. Such steps were applied to prepare tweets in objectives 2 and 3 for analysis. Figure 3.3 shows how a tweet, once translated to English, is preprocessed before analysis. These steps include cleaning the text by removing contractions, special characters, user names, and links. Contractions (e.g, "can't" or "won't”) are expanded to their complete forms (e.g. "cannot" and "will not"). Special characters and punctuations are often removed or replaced to focus on the essential words and content. User names, starting with "@" symbols, are typically removed to remove personal identifiers. Links and URLs are eliminated to avoid biasing the analysis towards external web resources. Stopwords are commonly occurring words (e.g, "the," "is," or "and") and are often removed as they contribute little to the overall meaning. Then, stemming is usually applied to reduce words to their base or root form (e.g. "running" and "runs" are converted to "run"). Lemmatization aims to transform words to their canonical form, considering the word's part of speech (e.g. "better" to "good").. Finally, tokenization aims to split the text into individual units or tokens, usually words or subwords, allowing for further analysis and processing at a granular level.


Tweet A
Diabetes is a chronic condition, find out more about it on my account.
Tweet B
I ate so much sugar that I now have diabetes!!!! #candies
Tweet C
I was diagnosed with #t1d in 2015 and I've always found insulin affordable. Good for me!!
Personal vs non-personal classifier
Jokes / irony classification
Preprocessing

Remove contraction/special words	Remove stopwords	Stemming	Lemmatization	Tokenization
I was diagnosed with type 1 diabetes in 2015 and I have always found insulin affordable. Good for me!!	diagnosed type 1 diabetes 2015 insulin affordable. Good	diagnos type 1 diabet 2015 insulin afford . Good	diagnos type 1 diabet 2015 insulin afford . good	diagnos type 1 diabet 2015 insulin afford good

Figure 3.3. Classification and preprocessing of diabetes-related tweets.

Sentiment and emotion analysis

For all collected tweets, sentiment and emotion analyses were performed to gain insights into users' overall mood and reactions to tweeting about diabetes.

Sentiment analysis refers to automatically determining the sentiment in a tweet. It involves analyzing the text to identify whether the sentiment is positive, negative or neutral. One popular approach to sentiment analysis on Twitter is the rule of the VADER (Valence Aware Dictionary and sEntiment Reasoner) tool [94], a rule-based sentiment analysis model designed to handle social media text. It uses a pre-defined lexicon of words and phrases with assigned sentiment scores to assess the sentiment of a given tweet. Thus, by considering contextual information, punctuation and capitalization, VADER provides an overall score for each tweet.

Emotion analysis investigates the emotional content of the text. It aims to identify and classify specific emotions within the text, such as happiness, sadness, anger and fear.

Emotion analysis captures a more nuanced understanding of the emotional states reflected in tweets. A large dataset of tweets labeled according to 4 emotions (joy, anger, sadness, fear) was constructed from open online labeled databases and manually labeled tweets linked to diabetes. This dataset was then used to train a classifier.

While sentiment analysis is helpful to understand the general sentiment of a large volume of tweets, emotion analysis provides a deeper understanding of the emotions associated with specific topics or events. Sentiment and emotion analysis play crucial roles in understanding public opinion, customer sentiment, and social media trends to better understand public perception and identify emerging patterns.

Word Embedding

Word embeddings are NLP techniques representing words or phrases as vectors in a high-dimensional space [95]. This is based on the hypothesis that words with similar contexts share semantic similarities. Thus, word embedding models capture and quantify the semantic and syntactic relationships between words by mapping them to vectors, allowing computer systems to interpret and understand language numerically.

Word2vec and GloVe are two traditional approaches using unsupervised learning algorithms to generate static representations of words [95,96]. These models were trained on large text corpora and analyze word co-occurrences to determine vector representations. The generated embeddings encode semantic similarity. This means that words that appear in similar contexts tend to have similar vector representations, allowing operations such as measuring cosine similarity between word vectors to determine semantic relatedness. increase. For example, in a well-trained Word2Vec model, words like 'king' and 'queen' have closer vectors compared to unrelated words like 'car' and 'dog'. It can measure word similarity or determine the relationship between words. However, traditional word embeddings treat all occurring words as the same entity and lack contextual awareness.

In contrast, contextual word embeddings are a recent significant advance in NLP: they integrate contextualization into word representations. These models, such as ELMo [97], GPT [98], and BERT [99], use deep learning techniques, especially transformers, to process words in the context of whole sentences or whole texts. By considering surrounding words and sentence structure, contextual word embeddings create word representations that dynamically adjust based on the context. This contextualization allows embeddings to capture nuances and clarify word meanings, removing the limitations of static embeddings. In the context of tweet analysis, these word embedding techniques were applied to convert individual tweets into vector representations. By representing tweets as vectors, it became possible to leverage clustering methods, enabling the grouping and categorizing of tweets based on their semantic similarities, facilitating further analysis and pattern recognition.

In our analysis, Word2Vec was used to transform each tweet into a 300-dimension vector by averaging its word vectors. Using cosine similarity as the distance metric, the semantic closeness of tweets was assessed using clustering methods.

Machine learning

Machine learning is a subfield of artificial intelligence as seen on Figure 3.4. It can be used to extract meaningful insights and to make predictions from data. It encompasses a range of algorithms and techniques that enable computer systems to learn patterns and relationships from data without being explicitly programmed.

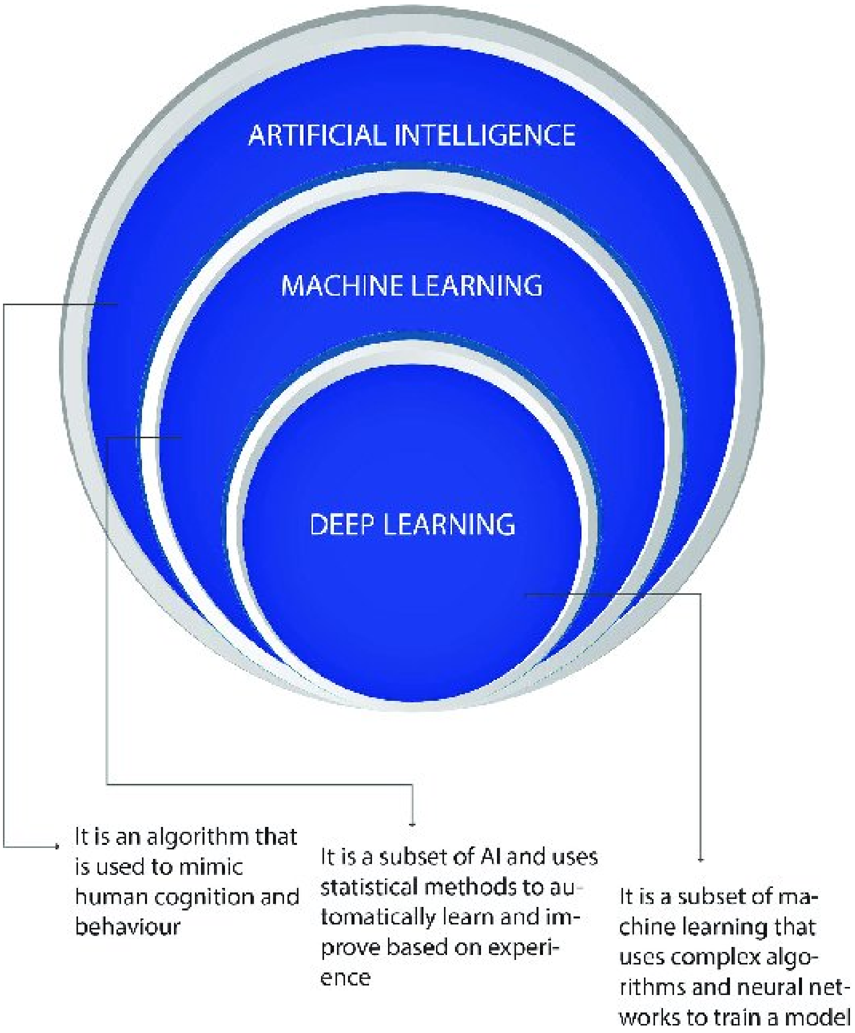

ARTIFICIAL INTELLIGENCE
MACHINE LEARNING
DEEP LEARNING
It is an algorithm that is used to mimic human cognition and behaviour
It is a subset of AI and uses statistical methods to automatically learn and improve based on experience
It is a subset of machine learning that uses complex algorithms and neural networks to train a model

Figure 3.4. Difference between Artificial Intelligence, Machine Learning and Deep Learning. From Khan, Protima, et al. "Machine learning and deep learning approaches for brain disease diagnosis: principles and recent advances." IEEE Access 9 (2021): 37622-37655.

Supervised machine learning involves training models on labeled datasets where the desired output or target variable is provided. These models learn to generalize from the labeled examples and make predictions on unseen data. They include algorithms such as decision trees, support vector machines, and neural networks. They are used for classification and regression tasks. The goal is to accurately predict outcomes based on input features and discover underlying patterns in the data. For instance, diabetes-related tweets were manually labeled according to the type of diabetes that was described in the tweet to create a classifier on users' type of diabetes. The notion of classification will be explained in the next section, under "Classification".

Unsupervised machine learning, on the other hand, deals with unlabeled data. The objective is to identify inherent patterns or structures within the data itself. Clustering and dimensionality reduction techniques are commonly used in unsupervised learning. Dimensionality reduction techniques reduce the number of features while preserving the information, making it easier to analyze and visualize high-dimensional data. Clustering algorithms group similar data points together, enabling the discovery of natural clusters or segments in the data.

There are several metrics commonly used to assess the efficiency and performance of a binary classification machine learning model [100]. Some of the typical metrics include:

- **True Positive**: data points that were correctly predicted as positive by the model.
- **False Positive**: data points that were incorrectly predicted as positive when they were actually negative.
- **True Negative**: data points that were correctly predicted as negative by the model.
- **False Negative**: data points that were incorrectly predicted as negative by the model.
- **Accuracy**: it measures the overall correctness of predictions by calculating the proportion of correctly classified instances.
- **Precision**: it represents the ratio of true positive predictions to the total number of positive predictions, representing the model's ability to avoid false positives.
- **Recall** (also called sensitivity): it measures the ratio of true positive predictions to the total number of actual positive instances, indicating the model's ability to capture all positive samples.
- **F1-score:** the harmonic mean of precision and recall, providing a balanced measure of a model's performance.

Classification

Classification is a task that involves predicting the class or the category of a given input based on its features or attributes [101]. It is a supervised learning approach where a model is trained using labeled data consisting of input samples and their corresponding class labels. Classification aims to learn a decision boundary or a mapping function that can accurately classify new, unseen instances into predefined classes. Classification techniques can be used in various domains, such as spam detection, sentiment analysis, medical diagnosis, image recognition, and fraud detection.

The input data for classification can have various forms, including numerical, categorical, or textual features. Before training the model, preprocessing steps such as data cleaning and normalization can be performed to enhance the quality and representativeness of the data. In the classification process, the model learns patterns and relationships from the input data to make predictions. It can be trained using a variety of algorithms, such as logistic regression, support vector machines, decision trees, random forests, or neural networks. These algorithms employ different strategies to optimize the decision boundary and minimize the classification error.

Transformers

Transformers are neural network architectures used to capture contextual relationships in language data by leveraging self-attention mechanisms [102]. They overcome the limitations of traditional recurrent neural networks by allowing parallel processing of words in a sequence. This parallelism allows Transformers to model long-range dependencies and capture complex linguistic patterns more effectively.

BERT

Particularly, BERT (Bidirectional Encoder Representations from Transformers) is one of the most prominent transformer-based models [99]. It is a pre-trained model that learns contextual word representations from vast amounts of unlabeled text. By training on diverse language data, BERT acquires a deep understanding of semantics, syntax, and contextual relationships between words. This knowledge is then transferred to new tasks by fine-tuning the pre-trained model on task-specific labeled data. For tweet analysis, a specialized variant of BERT called BERTweet has been developed [103]. BERTweet is designed to handle the unique characteristics of Twitter data, such as the presence of hashtags, user mentions, and emoticons. It is pre-trained on a large corpus of Twitter data, allowing it to capture the nuances and language patterns specific to tweets.

In this work, BERTweet was used to develop classifiers that can effectively differentiate personal tweets discussing personal experiences with diabetes from institutional or non-personal content. This classification task helps in identifying tweets that provide valuable insights into individuals' real-life experiences with diabetes. Furthermore, BERTweet was also used to filter out tweets containing jokes and irony related to diabetes. By using the contextual understanding and semantic knowledge captured by

BERTweet, tweets with humorous intent were distinguished and excluded from the analysis, ensuring the focus remained on genuine and informative content related to diabetes.

Sentence Transformers

Sentence Transformers represent an evolution and extension of the classic transformer architecture and models like BERT [104]. Sentence transformers specifically target the generation of sentence-level embeddings. Sentence Transformers are generating contextually aware embeddings and this excel in capturing semantic similarities and relationships between sentences. They go beyond simple averaging or pooling of word-level embeddings, effectively modeling sentence semantics and capturing the nuances of language at a higher level of abstraction. This makes sentence transformers highly effective in tasks such as paraphrase detection, document classification, information retrieval and semantic similarity assessment.

In the context of a Virtual Digital Cohort Study focused on diabetes, Sentence Transformers were used to compute the contextual similarities between tweets and specific concepts related to diabetes. This approach was used in order to handle synonyms. When computing the similarities between a tweet and a set of keywords related to a concept, such as “Mental Health” (as detailed in Appendix 6), the concept of mental health gathered keywords such as "depression, distress, burnout, anxiety, exhaustion, anxious, burden." By assessing the similarities between tweets and this set of terms, we could deduce if a tweet was associated with mental health or not, based on a predefined similarity threshold (0.41 in the case of the “Mental Health” concept as can be seen in Appendix 6). This method allowed us to identify discussions related to mental health even if the exact keywords weren't explicitly mentioned in the tweet. By gauging the similarity score, we were able to classify tweets' relevance to mental health topics while successfully addressing synonyms.

This strategy was used to emulate the functionality of the Linguistic Inquiry and Word Count (LIWC) [105]. When given a text, LIWC will be matching words to its internal dictionary in order to offer insights into the emotional, cognitive, and structural components present in the input texts. However, there are licensing fees associated with its use. This can be a hindrance to researchers and organizations with budgetary constraints, and it also would have posed a problem to replicating studies like VDCS if we used such an approach. Moreover, unlike the approach we employed with

Sentence-Transformers, LIWC does not account for the broader context in which words appear. This could lead to potential misinterpretations or overlooking nuanced meanings of tweets. Thus, using Sentence Transformers allows the Python package introduced in Chapter 6 to circumvent these limitations, enabling more nuanced and cost-effective text analysis.

Prediction

Prediction in machine learning is the process of estimating or forecasting a variable based on specific inputs. Prediction involves selecting the best algorithm (e.g, linear regression, decision trees, random forests, or neural networks) to capture and generalize the relationships between the input features and the target variable. Input data can be numerical, categorical, or textual features, and preprocessing steps like data cleaning, feature engineering, and normalization may be applied to enhance the quality and relevance of the input data. In this work, prediction techniques were used to estimate the rates of anger, sadness, joy, and fear in each tweet, as well as each users' gender and type of diabetes. By using machine learning algorithms and leveraging the input features derived from the textual content of the tweets, the model was able to analyze and predict the emotional states conveyed by the different users.

Clustering

Clustering is a technique in unsupervised machine learning that aims to group similar data points together based on their similarities [106]. It is a process of partitioning a dataset into clusters such that data points within the same cluster are more similar to each other compared to those in the other clusters. Clustering can help to uncover underlying patterns in the data, without any prior knowledge of the class labels.

K-means

In this thesis, K-means clustering algorithm [107,108]was used to analyze diabetes-related tweets to identify concerns and determinants of diabetes burden. This aims to gather tweets into distinct groups, ensuring that tweets within a cluster are more similar to each other than to those in other clusters. The algorithm workflow is the following:

1. Randomly select centroids of the K clusters from the dataset or specify initial centroid positions.
2. Assign each data point to the nearest centroid initialized in step 1 based on a distance metric, typically Euclidean or Cosine distance. This step forms initial clusters.
3. Recalculate the centroid of each cluster by taking the mean of all data points assigned to that cluster.
4. Repeat steps 2 and 3 until convergence, which occurs whether when the assignments and centroids no longer change significantly or when a predefined number of iterations is reached.

The effectiveness of K-means depends on several factors, including the choice of K (the number of clusters), the initialization strategy, and the distance metric used. Particularly, there are several methodologies to find the best number of clusters K. Two widely used approaches are the elbow method and the silhouette score.

The elbow method is a graphical technique that helps determine the optimal number of clusters by evaluating the trade-off between the number of clusters and the within-cluster sum of squared distances. The sum of squared distances is then plotted against different values of K. The "elbow" point is where the rate of decrease in inertia starts to level off. The idea is to select the value of K at the elbow point, as it indicates the number of clusters where adding more clusters does not provide significant improvement in clustering quality. Still, the elbow method does not always give a clear elbow point, particularly when the data is complex.

The silhouette score is an approach used to assess how well each data point fits into its assigned cluster compared to other clusters. The silhouette score ranges from -1 to 1. A value close to 1 means well-separated clusters while a value close to -1 suggests data points assigned to the wrong clusters. A value close to 0 indicates overlapping or ambiguous clusters. The silhouette score provides a more objective measure for choosing the optimal number of clusters, as it considers both cohesion within clusters and separation between clusters. The best number of clusters can be chosen by selecting the value of K that maximizes the average silhouette score across all data points.

In our work, to determine the optimal number of clusters of diabetes-related tweets, we used both these methods. Firstly, we applied the elbow method, which helped us identify a potential range of values for the optimal k. After narrowing down the range of k values, the silhouette scores were calculated to make a final decision. Thus, for each k within the identified interval, we calculated the silhouette score, which measures the compactness of clusters and the separation between them. By evaluating the silhouette scores, we were able to assess the quality and distinctiveness of the resulting clusters. By combining the elbow method and the silhouette score,the best number of clusters was easily determined for the analysis, ensuring that the selected value maximized the similarity within clusters while minimizing the dissimilarity between them.

Survival analysis

Survival analysis is a statistical method to analyze time-to-event data, particularly in the context of studying disease progression or mortality. It can be used to investigate factors that influence the duration until an event of interest occurs. Survival analysis gives insights into the risk factors and outcomes associated with various health conditions. The results can then be used to develop strategies for prevention, treatment, and improve public health interventions.

Cox proportional hazards models or Cox models are statistical models used in epidemiology and survival analysis to study the association between predictor variables and the time until an event occurs. Cox models are based on the proportional hazards assumption. It assumes that the risk (or hazard) of an event is proportional to the risk for other individuals at any given time, regardless of their risk at inclusion. This assumption allows for the estimation of hazard ratios, which quantify the relative risk associated with different predictor variables. In the Cox model, the baseline hazard function remains unspecified, meaning it does not make assumptions about the baseline risk over time. Instead, the focus is on estimating the effect of covariates on the hazard function. The model provides hazard ratio estimates, indicating the multiplicative effect of each predictor variable on the hazard.

Chapter 4

The use of social media for health research purposes: a scoping review

Paper published in Journal of Medical Health Research (DOI: 10.2196/25736)

My contribution to this Chapter: *Conceptualisation. Title, abstract, and full-text screening. Data extraction and data cleaning. Data visualization. Drafting the manuscript. Review and final approval of the manuscript.*

Abstract

Introduction

As social media are increasingly used worldwide, more and more scientists are relying on them for their health-related projects . But so far, social media features, methodologies and ethical issues are unclear with no overview of this relatively young field of research.

Objective

This scoping review aimed to provide an evidence map of the different uses of social media for health research purposes, their fields of applications and their analysis methods.

Methods

We followed the scoping review methodologies developed by Arksey and O'Malley and the Joanna Briggs Institute. After developing search strategies based on keywords (e.g., Social media, health research), comprehensive searches were conducted in Pubmed/MEDLINE and Web of Science databases. We limited the search strategies to documents written in English and published between 2005/01/01 and 2020/04/09. After removing duplicates, articles were screened at title/abstract and at full text level by two independent reviewers. One reviewer extracted data that were descriptively analyzed to map the available evidence.

Results

After screening 1,237 titles and abstracts and 407 full-texts, 268 unique articles were included, dating from 2009 to 2020 with an average annual growth rate of 32.71% for the 2009-2019 period. Studies mainly come from America (64.55%, N=173/268, including 151 from the USA). Articles used machine learning or data mining techniques (N=60/268) to analyze the data, discussed opportunities and limitations of the use of social media for research (N=59/268), assessed the feasibility of recruitment strategies (N=45/268) or discussed ethical issues (N=16/268). Communicable (e.g., influenza, N=122/268) and then chronic (e.g., cancer, N=40/268) diseases were the two main areas of interest.

Conclusions

Since their early days, social media have been recognized as a resource of high potential for health research purposes but yet the field is still suffering from a strong heterogeneity in the methodologies used, which prevents the research from comparison and generalisability. For the field to be fully recognized as a valid, complementary approach to more traditional health research study designs, there is now a need for more guidance by types of use of social media for health research, both from a methodological and an ethical perspective.

Keywords

Social media; Public health; Epidemiology; Research; Health; Medical; Social Networking; Infodemiology; eHealth; Text Mining; Natural Language Processing

Background

Social media (SM), sometimes confused with “social networks”, refer to new forms of media that involve interactions between users [109] in personal (e.g., Facebook) or more professional (e.g., Linkedin) ways. In 2010 in the US, 80% of adults used the internet to search for health-related information and 11% of SM users posted comments, queries or information about health or medical content [110]. Every user activity on the internet generates a unique digital footprint which can be collected for health research. [36]. However, SM are not only used in a personal way. Indeed, academics are also increasingly using SM to share their work and disseminate their findings [111].

Since the creation of SM in 2004-2005 and with 3.81 billions active social media users in April 2020 [112], concepts like infodemiology and infoveillance have emerged. The term “infodemiology” refers to the science of using the internet to improve public health, while “infoveillance” refers to the science of syndromic surveillance using the internet [113]. These opportunities have been seized through the years in order to create new methodologies for health research to cope with the issues raised by traditional methods (e.g, difficulty of recruitment [114]).

Previous works have already been published about the use of SM for health research. However, they were either focusing on a specific type of SM (e.g., blogs [115]), on a specific field of health research (e.g., children maltreatment [116]) or on a specific methodology (e.g., recruitment of study participants [117]). Other reviews discussed the overall use of SM for health research [118,119] but did not provide any insights on the analysis techniques nor the ethical issues. This is why we decided to perform a scoping review on the use of social media for health research purposes with the objective of not only detailing the opportunities and challenges that face this recent field of research in the age of digital health and artificial intelligence, but also providing details on the methodologies that are mainly applied by health researchers.

Review questions

The overall research questions were :

1. How have SM modified or complemented traditional health research?
2. What are the different fields of application of this approach?
3. What are the different methodologies for SM data analysis?

Methods

The objectives, inclusion criteria and methods for this scoping review were specified in advance and documented in a protocol [120]. This scoping review followed the methodological framework introduced by Arskey and O'Malley in 2005 [121] and the methodology manual published by the Joanna Briggs Institute for scoping reviews [122]. It is reported in accordance with the PRISMA Extension for Scoping Review (PRISMA-ScR) guidelines [123].

Search strategy

An initial literature search was first manually conducted to identify the health research fields in which SM are mostly used and developed. Then, the literature search was performed through PubMed/MEDLINE and Web of Science. The search strategy, highlighted in Table 1, included two sets of search terms: (i) one linked with SM (e.g., Social media) and (ii) research (e.g., Health research, Biomedical research). In order to capture the evolution of SM uses for health research over the years, databases were searched between 1 January 2005 and 9 April 2020. The term "Social network" was also searched as it is often misused as a synonym of SM. An additional list of 5 relevant articles [124–128] was manually searched to identify any other potentially relevant articles not yet captured. These articles were chosen in order to retrieve more articles about infodemiology, ethical issues or the use of SM data. A snowball searching technique was adopted with these 5 articles in which citations within articles were searched and kept if relevant to the review.

Eligibility criteria

This review was guided by the PCC (Population, Concept, Context) framework suggested by the Joanna Briggs Institute [129]. We did not have any restriction about the population, we took any relevant publications regardless of the age, the origin or the gender of the studied populations. The concept was the use of social media and the context was health research. The eligibility criteria were any journal article that described the use of social media platforms or social media data for health or medical research purposes. We excluded from our review, articles that were not directly related to research such as the use of social media among patients, patient associations or communities, organizations, healthcare professionals for their day-to-day practice. Documents related to the mining of social media data to detect prescription drug misuse and abuse were eligible for inclusion.

Grey literature and studies about non-human subjects were excluded as well. We included full-texts that reported on at least one of the following outcomes: (i) SM data analysis, (ii) recruitment through SM, (iii) methodology for SM research and (iv) ethical issues of using SM for health research. Only English-language articles were retained. The inclusion and exclusion criteria and the search string are summarised in Table 4.1.

Table 4.1 : Inclusion criteria, exclusion criteria and search strings

Inclusion criteria	- written in English - published between 2005/01/01 and 2020/04/09 - deals with the use of SM by researchers
Exclusion criteria	- not about health research - not related to social media (e.g., social network analysis) - not about human subjects - no relevant information (e.g., methodology) about the use of SM for health research - no relevant characteristics of SM
Search string in Pubmed	((("Social Media"[MH]) OR ("Social Media"[TW])) AND (("Biomedical research"[MH]) OR ("Medical research"[TW] OR "Biomedical research"[TW]) OR ("Health research"[TW] OR "Health services research"[TW]))) OR ((("Social networking"[MH]) OR ("Social network"[TW] OR "Social networks"[TW] OR "Social networking"[TW])) AND (("Biomedical research"[MH]) OR ("Medical research"[TW] OR "Biomedical research"[TW]) OR ("Health research"[TW] OR "Health services research"[TW]))) Filters: Journal Article; Publication date from 2005/01/01 to 2020/04/09; Humans; English
Search string in Web of Science	(TS="Social Media" OR TS= “Social networking” OR TS= “Social network” OR TS= “Social networks”) AND (TS="Biomedical research" OR TS="Medical research" OR TS="Health research" OR TS="Health services research") AND (PY=(2005-2020)) AND (LANGUAGE: (English)) Indexes=SCI-EXPANDED, SSCI, A&HCI, CPCI-S, CPCI-SSH, ESCI Timespan=All years

Study selection process

A two-step screening was performed after duplicates removal. First, titles and abstracts were screened in order to define the eligibility of each article. Publications with title or abstract not meeting the eligibility criteria were excluded. Then, the full texts having passed the first step were screened and only articles meeting the eligibility criteria were kept. All screening levels were conducted with CADIMA [130], a free web tool to facilitate the conduct and the documentation of literature reviews [131]. Two reviewers screened articles (GF, CB) independently and consistency checks were performed thanks to CADIMA.

Data extraction

Data were abstracted on : (i) the country of origin, (ii) the aims of the study (e.g., to map ethical issues when using SM for health research), (iii) the type of study (e.g., recruitment feasibility assessment), (iv) the research field (e.g., mental health research) (v) the studied population (e.g., adolescents), (vi) the type of SM (e.g., Facebook), (v) the methodology (e.g., paid advertisement), (vi) the outcomes of the study (e.g., efficiency of recruitment via SM) and (vii) the key findings for our scoping review (e.g., possibility to recruit on SM). Data were extracted and cleaned by a first reviewer (CB), then verified and approved by a second reviewer (GF).

Methodological quality appraisal

Because this is a scoping review, we did not appraise methodological quality or risk of bias of the included articles.

Analysis and presentation of results

We conducted a descriptive analysis of the characteristics of the included literature. We described the included articles according to the journal of publication, publication date, country of origin (location of the corresponding author), altmetric score (automatically calculated weighted count of all of the attention a research output has received) [132], type of SM, type of population and type of disease studied. We categorised the diseases in 7 categories : (i) Chronic diseases (e.g., diabetes), (ii) Communicable diseases (e.g., influenza), (iii) Alcohol/Smoking (e.g., vaping), (iv) Mental health (e.g., depression), (v) Lifestyle (e.g., nutrition outcomes), (vi) Drug/Medication (e.g., drug use disorder) and (vii) Other (e.g., child maltreatment). Descriptive statistics and corresponding plots were computed (N, means, frequencies) with R (version 1.1.463).

Results

Search results

The initial search conducted in April 2020 revealed 1,343 results. An additional 96 articles were retrieved through a snowballing technique based on 5 relevant articles (16-20). This resulted in a total of 1,439 articles and duplicates (N=202) were removed. Then, 1,237 titles and abstracts were screened which led to the exclusion of 830 articles. Overall, 407 studies were included to screen as full-text papers, of which 139 were excluded. The main reasons for exclusion were the study (i) did not contain relevant characteristics of SM for health research (N=25), (ii) did not relate to SM (N=45) or (iii) was not about health research (N=33). 268 studies were included in the analyses. Figure 4.1 shows the flow diagram of the articles selection. Lastly, Multimedia Appendix 1 displays the characteristics of the 268 included studies (author(s), year of publication, country, title, aim of the study, type of social media, studied population and disease).

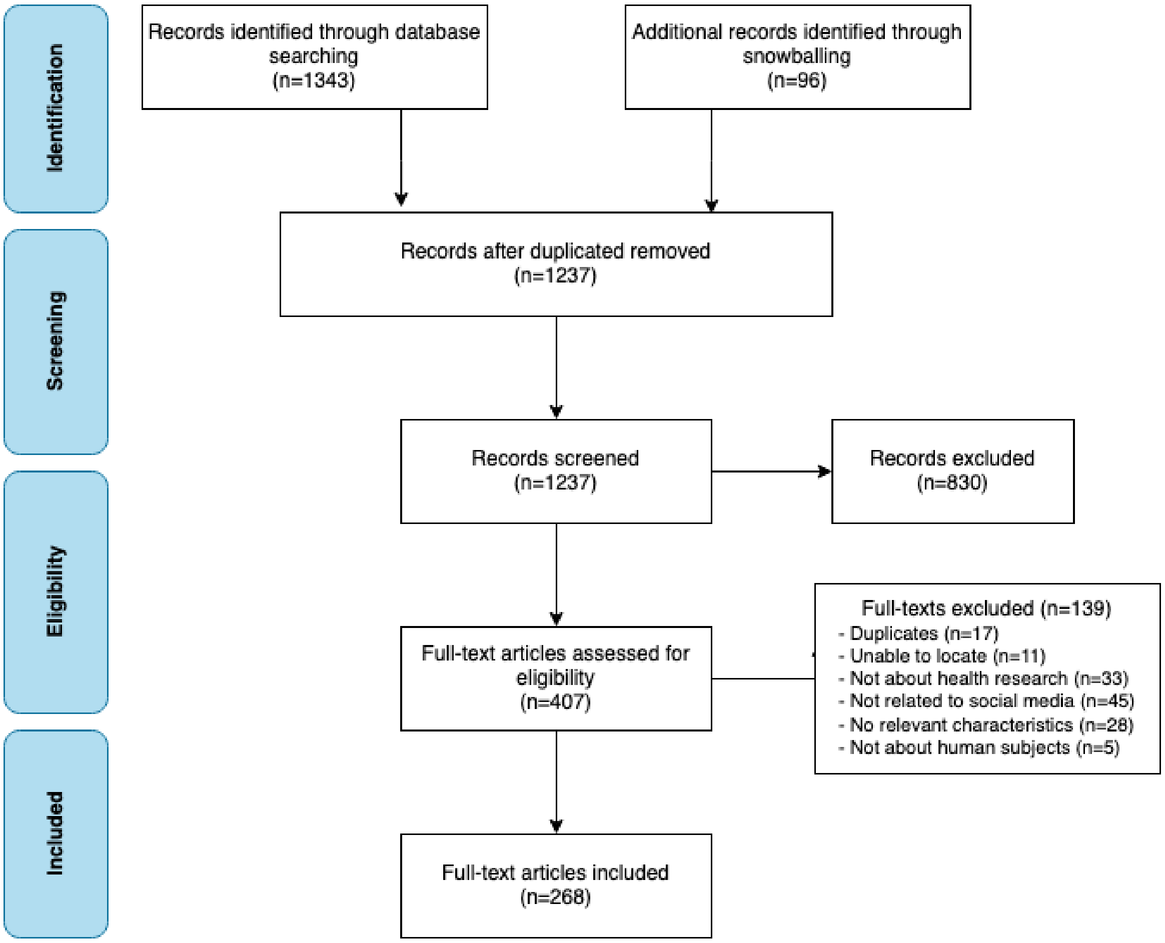

Identification
Screening
Eligibility
Included
Records identified through database searching (n=1343)
Additional records identified through snowballing (n=96)
Records after duplicated removed (n=1237)
Records screened (n=1237)
Records excluded (n=830)
Full-text articles assessed for eligibility (n=407)
Full-texts excluded (n=139)
- Duplicates (n=17)
- Unable to locate (n=11)
- Not about health research (n=33)
- Not related to social media (n=45)
- No relevant characteristics (n=28)
- Not about human subjects (n=5)
Full-text articles included (n=268)

Figure 4.1: Flow diagram of the included studies

Distribution of studies

In all, we included 268 unique records from 155 different journals. Table 4.2 displays the 10 most common journals in which the included studies were published : 20.5% of articles were published in the Journal of Medical Internet Research or sister journals JMIR Research Protocols and JMIR Public Health and Surveillance. PLoS ONE is the second most common journal with 3.7% of articles.

Table 4.2 : Top 10 most common journals publishing work using social media for health-research purposes

Name of the journal	Number of articles	Corresponding %
Journal of Medical Internet Research	39	14.55

PLoS ONE	10	3.73
JMIR Research Protocols	9	3.36
JMIR Public Health and Surveillance	7	2.61
American Journal of Public Health	5	1.87
The American Journal of Bioethics	5	1.87
BMC Medical Informatics and Decision Making	4	1.49
International Journal of Environmental Research and Public Health	4	1.49
PLoS Computational Biology	4	1.49
Digital Health	3	1.12

A total of 1,025 authors took part in the writing of the included studies. Figure 1 (Multimedia Appendix 2) provides the co-authorship network of all these authors. The largest set of connected authors includes 57 authors and is shown in Figure 2 (Multimedia Appendix 2).

Even though our research date range was from 2005 to 2020, no remaining articles from the 268 are dated before 2009. In Figure 4.2, it can be seen that the number of publications is growing through the years corresponding to an average annual growth rate of 32.7% for the 2009-2019 period. This suggests that the field of health research supplemented by SM has gained interest for the last 11 years. Earlier studies concentrated more on the use of SM for health research in general and the opportunities for the study of communicable diseases. Most recent studies include more frequently recruitment strategies and methodologies. Figure 4.3 displays the distribution of articles respectively to the continent of publication. Most articles are from America (64.6%, N=173/268, including 151 / 87.3% from the USA), 18.7% are from Europe (N=50/268), 11,6% are from Oceania (N=31/268), 4.9% are from Asia (N=13/268) and 0.2% from Africa (N=1/268).

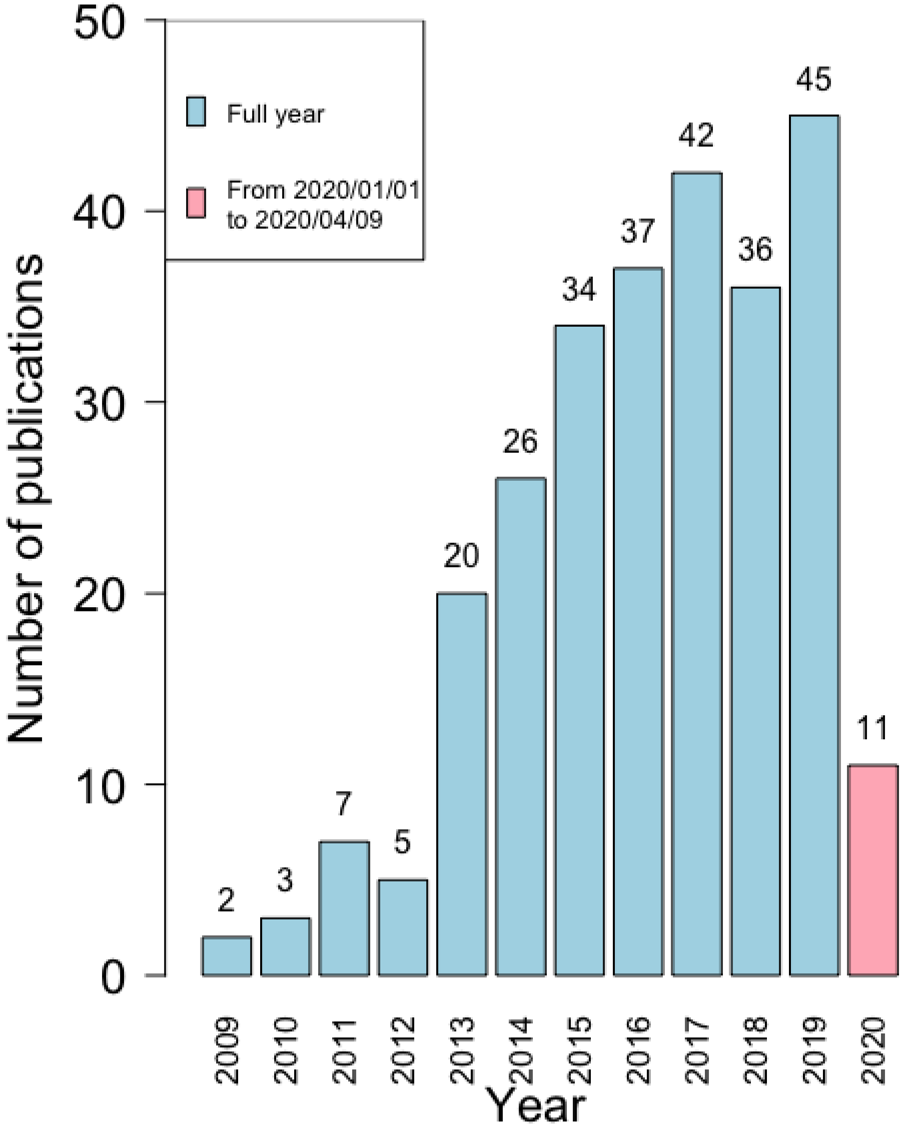


Figure 4.2: Distribution of publications by year of publication

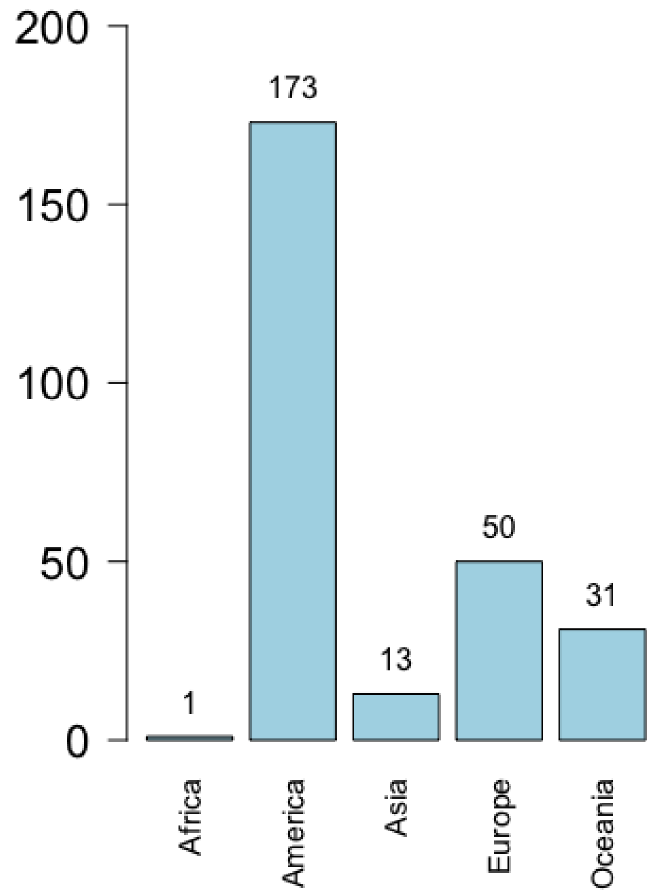


Figure 4.3: Distribution of publications by geographic location (as assessed by the location of the corresponding author)

Social media

Among all the retrieved articles, 57,9% (N=155/268) used or described at least one specific SM. From these articles, as can be seen in Figure 4.4, 42.6% (N=66/155) are based on Twitter, 34.2% (N=53/155) on Facebook and 11% (N=17/155) on several SM (e.g., combining Facebook, Instagram and Snapchat [133]). The remaining 12.3% (N=19/155) are distributed within Instagram, Reddit, forums, blogs, Weibo and Youtube.

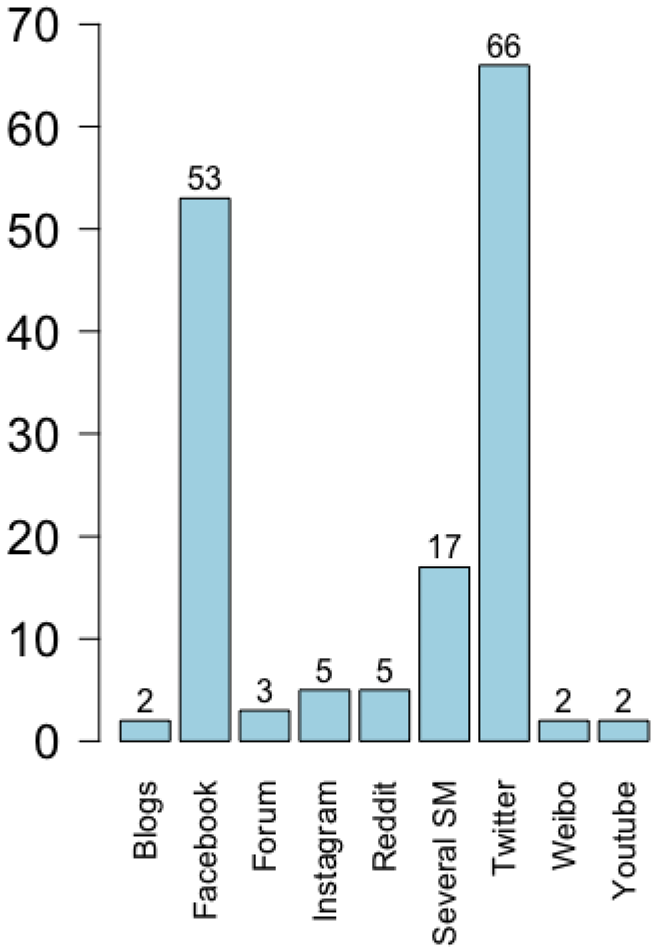


Figure 4.4 : Distribution of publications by social media (N=155)

Focused populations

A total of 80.2% (N=215/268) of included articles was not focusing on any specific population. In articles that studied a specific sub-population (N=53), youth is the most common one (64.2%, N=34/53), followed by women (13.2%, N=7/53), families (9.43%, N=5/53), men (1.9%, N=1/53) and other (11.3%, N=6/53), as shown in Figure 4.5. The “Other” category gathers adults (N=2/6), Chinese migrants (N=1/6), elderly people (N=1/6), emergency nurses (N=1/6) and researchers (N=1/6).

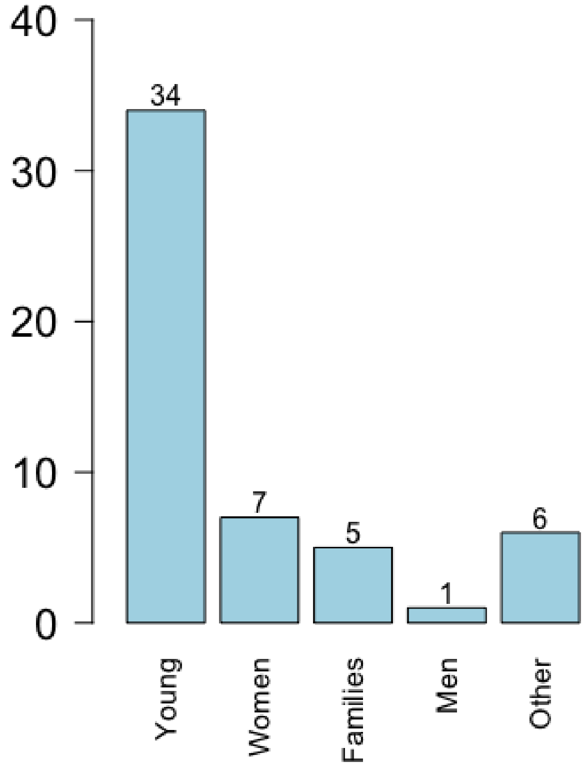


Figure 4.5: Distribution of publications per studied population (N=53)

Domain of health research

Besides, 45.5% (N=122/268) of publications dealt with a specific disease or condition. Indeed, as shown in Figure 4.6, 32.8% (N=40/122) of articles studied communicable diseases, 19.7% (N=24/122) chronic diseases, 15.6% (N=19/122) with lifestyle (e.g., nutrition outcomes), 14.8% (N=18/122) with other conditions (e.g., drug use disorder), 9.8% (N=12/122) with alcohol/smoking (e.g., vaping) and 7.4% (N=9/122) about mental health (e.g., depression).

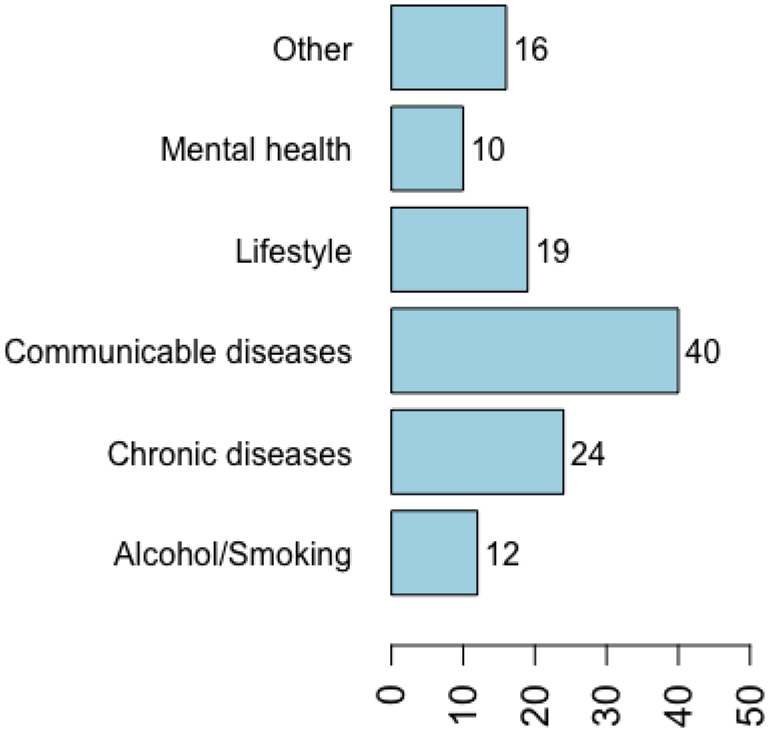


Figure 4.6: Distribution of publications by studied disease type (N=122)

Communicable diseases

In articles that discussed communicable diseases, influenza is the primary studied disease (45.0%, N=18/40), followed by HIV (20.0%, N=8/40) and HPV (7.5%, N=3/40), as shown in Figure 4.7.

Chronic diseases

In articles that discussed chronic diseases, a quarter studied cancer (25.0%, N=6/24), followed by diabetes (20.8%, N=5/24), cardiovascular diseases (e.g., congenital heart disease, 12.5%, N=3/24) and obesity (8.3%, N=2/34), as shown in Figure 4.8.

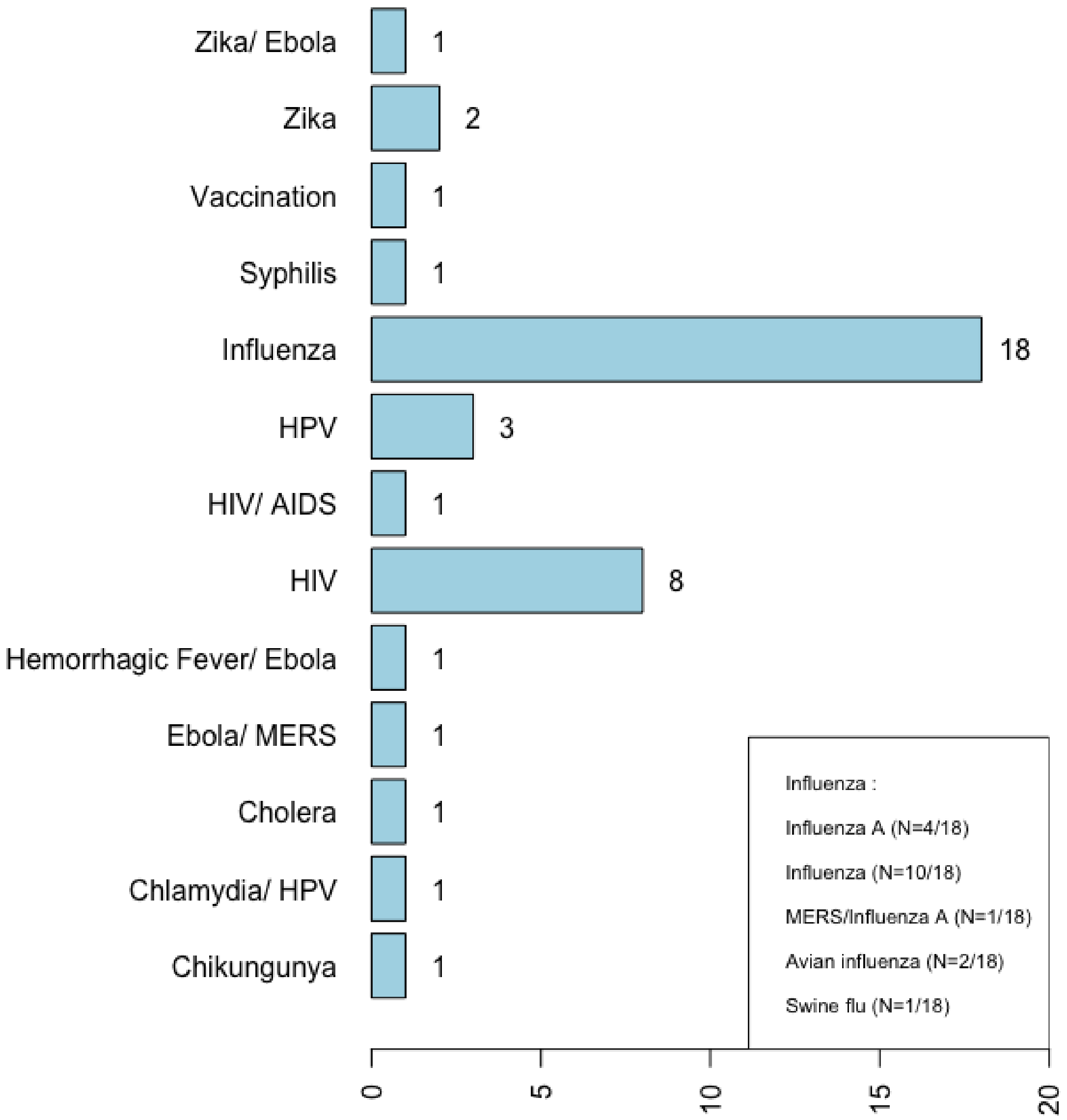


Figure 4.7: Distribution of publications in the communicable diseases category (N=40)

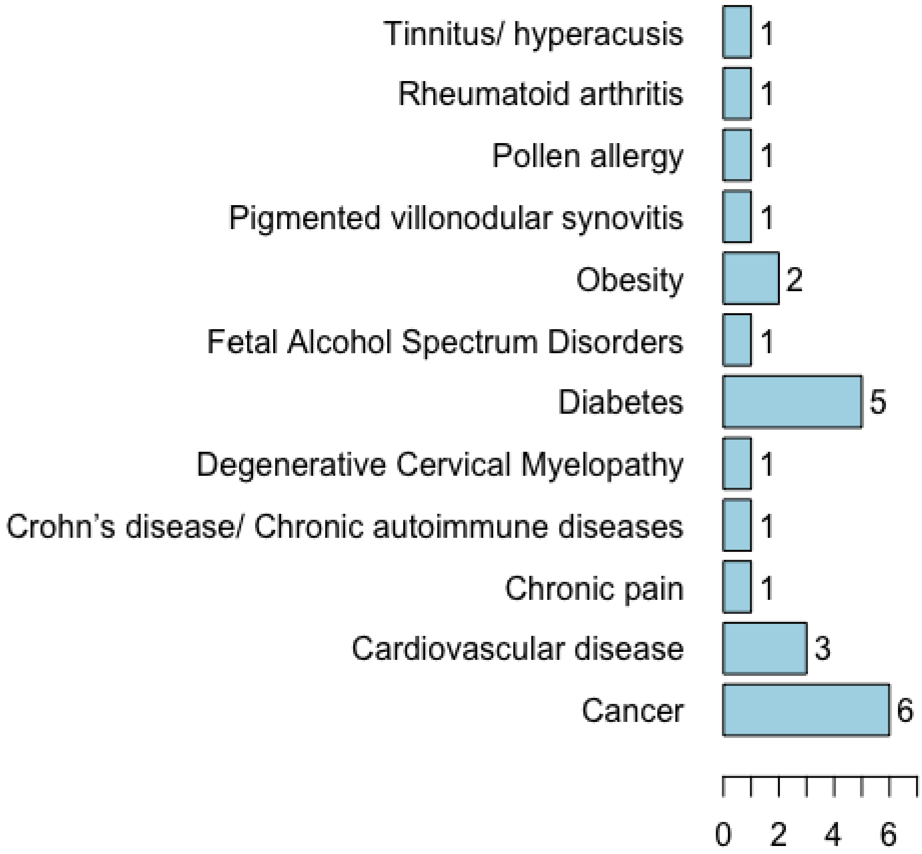


Figure 4.8: Distribution of publications in the chronic diseases category (N=24)

Dissemination

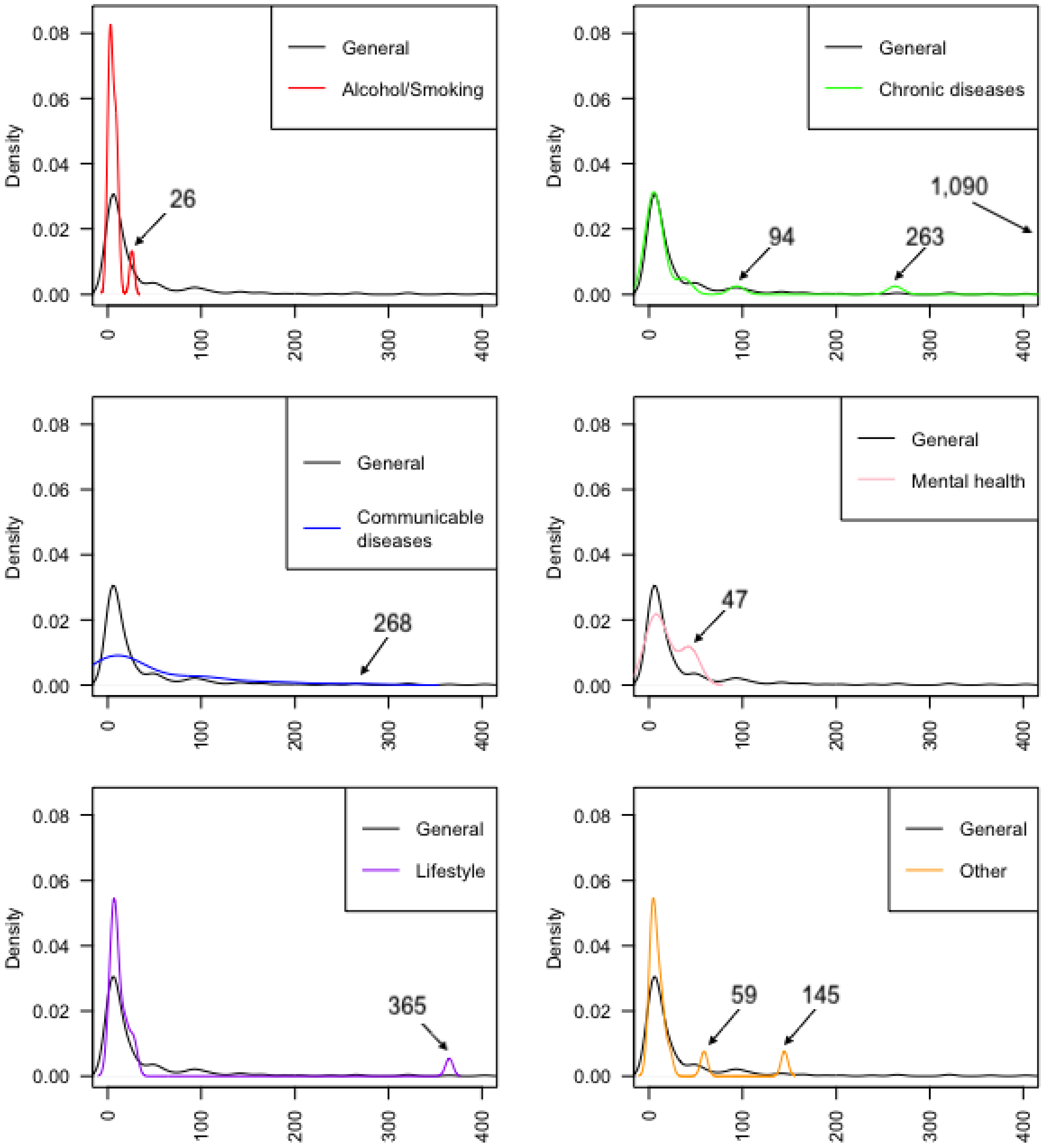


Figure 4.9: Distribution of altmetric scores by health research area

Corresponding publications to the pointed altmetric scores. Alcohol/Smoking : 26 [134]; Chronic diseases : 94 [27], 263 [135], 1,090 [136]; Communicable diseases : 268 [137]; Mental health : 47 [138]; Lifestyle : 365 [139]; Other 59 [140], 145 [141]. General corresponds to the altmetric scores of all studies.

Type of studies

Among all included studies, 22.4% (N=60/268) described the use of machine learning and data mining techniques 22% (N=59/268) discussed the opportunities and limitations of the use of SM for research, 16.8% (N=45/268) assessed the feasibility of recruitment strategies on SM, 6% (N=16/268) discussed the ethical issues when using SM for health research, 5.2% (N=14/268) gave methodologies for health research and 4.9% (N=13/268) illustrated the use of SM for dissemination. Guidelines for recruitment (3.4%, N=9/268), interventions of prevention (2.2%, N=6/268), crowdfunding (1.5%, N=4/268), sentiment analysis (1.5%, N=4/268), data anonymization (0.7%, N=2/268) and crowdsourcing (0.5%, N=2/268) are also considered.

Machine learning and other techniques

Machine learning techniques include text mining (28.3%, N=17/60), natural language processing (25%, N=15/60), data mining (20%, N=12/60), classification (16.7%, N=10/60), topic modelling (6.7%, N=4/60), deep learning (1.7%, N=1/60) and social network analysis (1.7%, N=1/60). In particular, support vector machine (28.3%, N=17/60), logistic regression (18.3%, N=11/60), latent dirichlet allocation (8.3%, N=5/60), convolutional neural network (8.3%, N=5/60), random forests (6.7%, N=4/60), decision trees (6.7%, N=4/60) and n-grams (6.7%, N=4/60) are the most used models. Stacked linear regression, bayesian network algorithm, non-negative matrix factorization, stochastic gradient, learning vector quantization and recurrent neural networks represent 1.7% (N=1/60) each. Lastly, these techniques are mostly used for data coming from Twitter (63.3%, N=38/60) and Reddit (5%, N=3/60).

Recruitment strategies

Studies assessing recruitment strategies feasibility applied paid advertisement (80%, N=36/45), free advertisement (e.g., posting in relevant Facebook groups [142]) (13.3%, N=6/45) and the combination of both advertisements (6.7%, N=3/45). Paid recruitment strategies include designing the ad, targeting the right audience with FB ad manager and measuring the impacts with FB analytics [143]. Moreover, 64.4% (N=29/45) of studies considered SM recruitment as effective (time-effective and efficient to recruit populations). Paid advertisement is evaluated as cost-effective in 83.3% (N=30/36) of studies and too costly in 5.6% (N=2/36). We found out that 80% (N=36/45) of recruitment are carried out on Facebook, 8.9% (N=4/45) on both Facebook and Twitter and 8.9% (N=4/45) on more than two SM (e.g., Facebook, Twitter, Craigslist, Tumblr, Linkedin, [144]). Lastly, a third of recruitment strategies included providing incentives to participants (e.g., gift cards).

Ethical issues

The main ethical issues raised are getting consent of online users (93.8%, N=15/16), protecting the privacy of users (in 87.5%, N=14/16), preserving confidentiality (56.3%, N=9/16), potential harms to participants (in 56.3%, N=9/16), preserving of anonymity (50%, N=8/16), data securing (in 43.8%, N=7/16), transparency of the research (43.8%, N=7/16), application of guidelines (in 43.8%, N=7/16), representativeness and self-selection bias (31.3%, N=5/16) and the risk of double accounts (12.5%, N=2/16).

Discussion

The overarching aim of this review was to scope the literature for evidence on the use of SM for health research. We were able to include 268 studies. Most of the included articles in this scoping review are dated from 2013 onwards, which is consistent with the worldwide growth of SM use over the last decade [145]. We identified three main SM used for health research : Twitter, Facebook and Instagram, the most popular platforms in 2020 [146]. The most studied populations are young adults and adolescents. This could be related to the elevated proportion of young people active on SM. In 2018 in the US, 51% of teens were on Facebook, 69% on Snapchat, 72% on Instagram and 85% on Youtube, thus SM seems to have great potential to focus on the young generations [147]. The majority of the included works focused on both communicable and chronic diseases.

The fields of application of SM in health research are broad. First, SM can be used to complement traditional methods. Traditional procedures can meet several limitations. When recruiting a specific population, traditional methods (e.g., fliers, advertising) can be expensive or limited by reach [114,148–150]. Complementing it with SM advertisements can cope with these limitations. Second, SM alone show high potential. Studies have concluded that SM paid advertisements can be an efficient and cost-effective tool to recruit [151–156]. SM, and particularly Facebook, appear not only to facilitate and complement traditional recruitment strategies to reach specific populations but to be efficient as well when used alone [152,157,158], specially to reduce time constraints or to target a large population [159]. Traditional disease surveillance, population surveillance and epidemiology methodologies can be improved by SM [126,150,160]. Pharmacovigilance and the detection of adverse drug reactions on SM proved to be efficient and to reduce time between the online report of an incident and its discovery.

[161–163] As SM users are increasing, generated data, or “Big data” is expanding. Such data can be collected and studied to improve disease and public health surveillance [164–166] to forecast diseases [167] or to improve research in a medical field [168]. Along with big data growth, machine learning and data mining techniques such as text mining and natural language processing are constantly evolving and are, thus, increasingly used in the field of public health research based on SM [169–171]. Twitter is mainly used for such work because Twitter developed a streaming application programmer’s interface (API). This is a free application that allows easy access to 1% of all Twitter data in real time, filtered by specific criteria (e.g., keywords) [172,173]. Lastly, SM can be directly used by health researchers to support prevention interventions to raise awareness and engage populations [174] and to crowdfund by promoting their researches on SM. Indeed, crowdfunding can be eased by establishing professional contacts through SM and sharing campaigns [175].

Still, the use of SM features and SM data for health research induces several ethical issues and limitations. Online data, like in Twitter for instance, is often considered as public and user consent is not provided for collecting it. Moreover, ensuring privacy protection of a data set when anyone has access to vast amounts of public information is difficult because data could be re-identified [176,177]. Safety features should be used to protect participants’ personal and sensitive information [125] or to protect participants from dangerous content posted by detractors or social media trolls (people who purposely provoke other SM users) [178]. These kinds of behaviors can also be oriented to researchers themselves and demotivate them. Moreover, data can represent only certain participants characteristics due to researchers self-selection or to coverage issues of underserved populations or minority groups who are disproportionately absent online. This can bias the representativeness of the sample and consequently, bias the findings and prevent from any generalisability [179,180]. When recruiting and providing incentives, users might be tempted to participate multiple times. Researchers should ensure that the study allows only one response from a given IP address [181,182]. A few guidelines and frameworks have already been created to guide health researchers in using social media and prevent such issues [183–187].

Strengths and limitations of this scoping review

The present work used a rigorous scoping review methodology from the manual by the Joanna Briggs Institute [188] throughout the entire process. It was guided by a previously published protocol [120]. To ensure a broad search of the literature, the search strategy included two electronic bibliographic databases and the snowball technique. There are some limitations to our scoping review process. We may not have identified all relevant articles in the published literature despite attempts to be as comprehensive as possible. We limited our review to documents written in English which may have led to missed relevant studies. Data were abstracted by one reviewer and verified by a second reviewer because of the important number of included publications.

Conclusion and recommendations

Our findings suggest that SM hold high potential to improve and complement existing health research studies. Indeed, some SM features can complement traditional research strategies and SM growing amounts of data hold great opportunities in the evolution of infoveillance and infodemiology. For researchers, SM can be an effective tool at almost every step of a study: from the development, the ideation, the recruitment and crowdsourcing to the dissemination of findings. Researchers should determine which SM is better fitting their objectives, as Facebook might be better for recruitment and Twitter for data collection in order to gain time and efficiency. Last but not least, we have observed a strong heterogeneity in the approaches used. We therefore recommend to take the existing guidelines into account and carefully think about the different ethical issues highlighted in this work before using SM for research.

Take-home messages from Chapter 4

1. The use of SM for health research has gained significant interest over the past decade, with a growing number of publications each year since 2009.
2. Twitter, Facebook and Instagram are the most commonly used SM platforms for health research, with Twitter being the most prevalent.
3. Young adults and adolescents are the most studied population on SM, likely due to their high engagement on these platforms.
4. The majority of studies focus on both communicable and chronic diseases, with a particular emphasis on influenza, cancer, diabetes and cardiovascular diseases.
5. SM offers opportunities to complement traditional research methods and recruitment strategies, with paid advertisements on platforms like Facebook proving to be cost-effective and efficient.
6. Machine learning and data mining techniques, such as text mining and NLP, are increasingly used to analyze SM data for health research purposes.
7. SM can enhance disease surveillance, public health research, and epidemiology by providing real-time data and improving the timeliness of detecting adverse drug reactions.
8. Ethical considerations include obtaining consent from online users, protecting privacy and confidentiality, addressing potential harms to participants, and ensuring data security.
9. SM can also be used for prevention interventions, raising awareness, and crowdfunding for health research projects.
10. Despite the opportunities, researchers should be aware of the limitations and challenges associated with using SM for health research, such as representativeness and self-selection bias.

Link with the following chapter

Building upon the findings of chapter 4, chapter 5 explores into **understanding the determinants of diabetes burden** through the analysis of diabetes-related tweets. By leveraging the rich information contained within social media data, this chapter aims to uncover additional insights and patterns that may not have been captured in traditional research approaches. The analysis of tweets from around the world will provide a broader understanding of the cultural, social, and behavioral factors influencing the burden of diabetes.

Chapter 5

Global diabetes burden: analysis of regional differences to improve diabetes care

Published in BMJ Open Diabetes Research and Care (DOI:10.1136/bmjdrc-2022-003040)

My contribution to this Chapter: *Conceptualisation. Data collection. Data labeling and classification. Data analysis. Data interpretation. Drafting the article. Article revision.*

Abstract

Introduction

The current evaluation processes of the burden of diabetes are incomplete and subject to bias. This study aimed to identify regional differences in the diabetes burden on a universal level from the perspective of people with diabetes.

Research Design and Methods

We developed a worldwide online diabetes observatory based on 34 million diabetes-related tweets from 172 countries covering 41 languages, spanning from 2017 to 2021. After translating all tweets to English, we used machine learning algorithms to remove institutional tweets and jokes, geolocate users, identify topics of interest and quantify associated sentiments and emotions across the 7 World Bank Regions.

Results

We identified four topics of interest for people with diabetes (PWD) in the Middle East and North Africa and another 18 topics in North America. Topics related to glycemic control and food are shared among six regions of the world. These topics were mainly associated with sadness (35% and 39% on average compared to levels of sadness in other topics). We also revealed several region-specific concerns (e.g., insulin pricing in North America or the burden of daily diabetes management in Europe and Central Asia).

Conclusions

The needs and concerns of PWD vary significantly worldwide and the burden of diabetes is perceived differently. Our results will support better integration of these regional differences into diabetes programs to improve patient-centric diabetes research and care, focused on the most relevant concerns to enhance personalized medicine and self-management of PWD.

What is already known on this topic

Twitter data can be a useful resource to monitor key concerns of people with diabetes, complementary to what can be achieved with questionnaires in clinical studies.

What this study adds

This study included a worldwide analysis of a dataset of 34 millions of Tweets from 172 countries to detect the most important topics of interest of people with diabetes and to study their differences across the seven World Bank Regions.

We have identified universal topics of concern. The concerns related to glycemic control and food are common to 7 and 6 regions of the world, respectively.

Other topics were found to be more important in some specific regions, such as insulin pricing in North America or the burden of daily diabetes management in Europe and Central Asia.

How this study might affect research, practice and/or policy

Our results can support the development of tailored diabetes programs at the regional level to focus on the most important concerns, and thus to enhance personalized medicine and self-management of people with diabetes.

Introduction

The term "burden of disease" describes the overall consequences (loss of health, social aspects, costs to society, death) caused by diseases, injuries, and risk factors worldwide and is often measured using Quality-Adjusted Life Years (QALYs) or Disability-Adjusted Life Years (DALYs).[189–191] However, QALYs and DALYs prevent us from understanding the drivers of the diabetes burden, such as the role of diabetes distress or the quality of care. Diabetes distress defines the emotional distress linked to living with diabetes and day-to-day management but also worrying about complications.[86] It has been shown that one in four people with type 1 diabetes and one in five people with type 2 diabetes have high levels of diabetes distress.[192] Emotional distress is associated with diabetes self-management and glycemic control issues.[82]

Conceiving patient-centered instruments helped measure the quality of care for PWD. Many of these have additional subscales ortheir evaluation aspects overlap.[82] These gaps in the assessment methods of the quality of care for PWD need to be identified. The most important factors must be prioritized and become objectives to address. As priorities for a person with diabetes in the USA may differ vastly between a PWD in Western Europe, the Middle East, or South Asia, determining the regional objectives is necessary to improve the lives of PWD. It is crucial to understand the regional differences in how the diabetes burden is perceived to integrate them into future diabetes programs. These could then address the most relevant local factors of diabetes burden.

One international source of data that captures the viewpoint of people with diabetes is Twitter. With more than 130 million users in 2019, it proved to be compatible with health research in various ways, but mainly to collect a considerable volume of data for public health surveillance, early event detection, outbreak prediction, and analysis of a population's sentiments and emotions.[43,193–197] Sentiment analysis aims to recognize polarity in texts (positivity, negativity, or neutrality), while emotion analysis determines the emotional state of an individual (anger, fear). Several diabetes communities have developed on Twitter, where users can share their experiences, ask for advice, or chat. They can be found with relevant hashtags (#dsma: Diabetes Social Media Advocacy, #gbdoc: UK Diabetes Online Community). It is thus possible to access large quantities of diabetes-related data from individuals and communities of PWD on Twitter. Social media data enables a better understanding of the principal daily concerns and associated emotions related to diabetes, diabetes management, diabetes distress, or diabetes burden.[198] More broadly, social media data may provide insights into how concerns differ between countries. As Twitter is embraced globally by numerous people and does not rely on predefined questions like evaluation scales, collecting and analyzing tweets can be considered an innovative and complementary way to understand PWD's feelings and concerns about their diabetes. *Ahne et al.* previously showed that such analysis could be efficient in identifying primary concerns in the USA.[88]

Because precision health starts by contextualizing the needs of the patients, we have tested the hypothesis that it is feasible to use a reproducible approach to analyze online data to better understand the determinants of diabetes burden and to identify regional differences that will serve to design more patient-centered diabetes programs in the future.[43]

Research Design and Methods

Data collection

Tweets are public by default and can be collected using the Twitter Application Programming Interface (API), which provides access to 1% of all Twitter data in real-time based on keywords. To collect diabetes-related tweets, we defined a list of 272 diabetes-related keywords such as *diabetes, insulin,* and *blood glucose* in 30 different languages (Appendix 3). Overall, the collection includes 34 million tweets published between May 2017 and April 2021. The data collected for this study only includes publicly posted tweets.

Preprocessing

The first step consisted of deleting duplicates and retweets to keep unique tweets and quote retweets (a retweet with an added comment). Secondly, non-English tweets were translated into English. Third, two classifiers were applied to keep only tweets with personal, non-joke, or non-ironic content from users sharing diabetes-related information about themselves or relatives. The workflow can be seen in Figure 5.1.

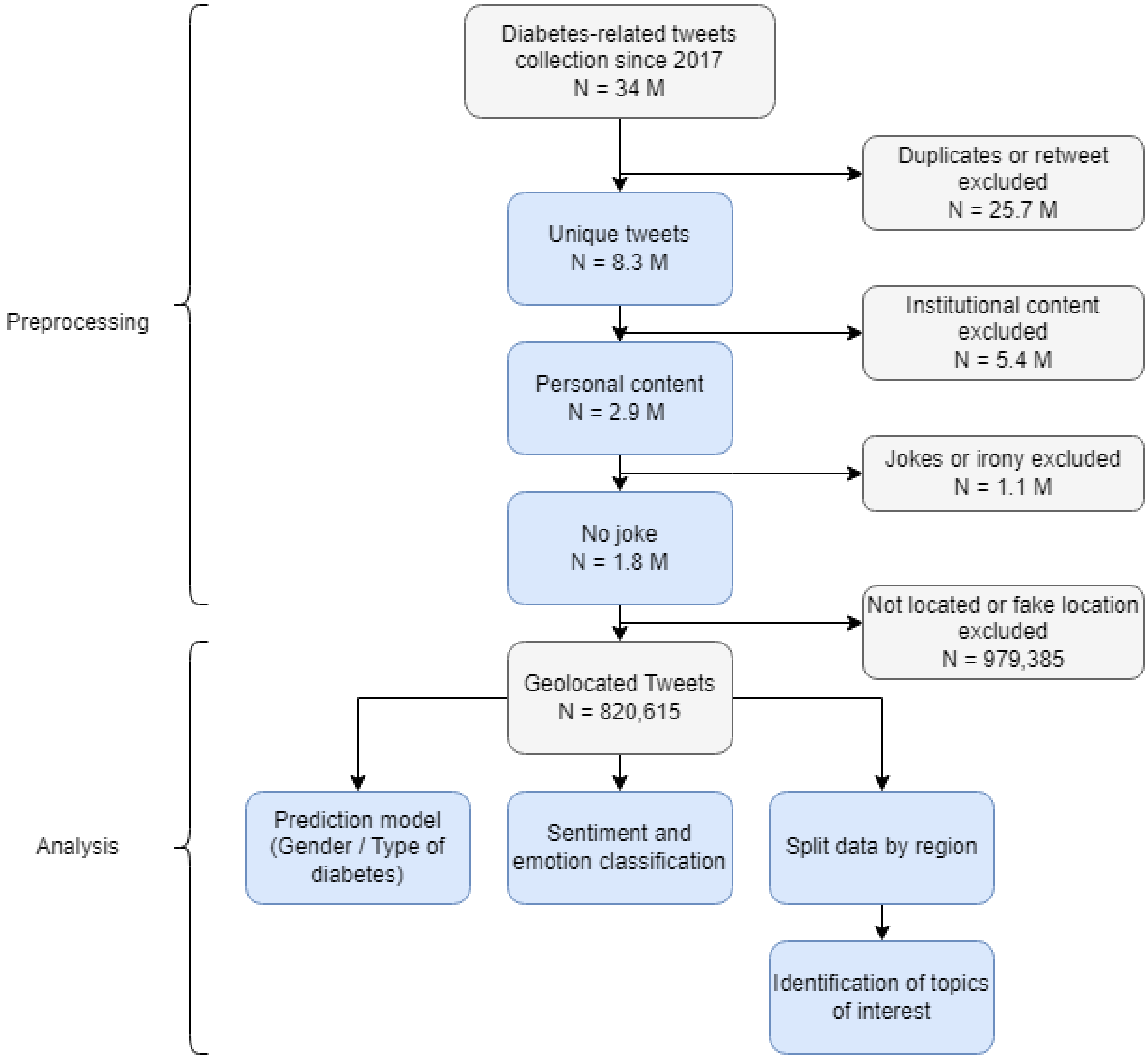

Preprocessing
Analysis
Diabetes-related tweets collection since 2017
N = 34 M
Duplicates or retweet excluded
N = 25.7 M
Unique tweets
N = 8.3 M
Institutional content excluded
N = 5.4 M
Personal content
N = 2.9 M
Jokes or irony excluded
N = 1.1 M
No joke
N = 1.8 M
Not located or fake location excluded
N = 979,385
Geolocated Tweets
N = 820,615
Prediction model (Gender / Type of diabetes)
Sentiment and emotion classification
Split data by region
Identification of topics of interest

Figure 5.1: Workflow showing the data preprocessing and analysis. Blue boxes correspond to steps where machine learning methods apply.

Geolocation

A tweet object provides meta-data, including information about the user account and location. The users provide their geographical area via an entry in their public profile. The precision in their description may vary. After applying the process described in Appendix 4, tweets were separated into the following seven regions: North America, East Asia and Pacific, Europe and Central Asia, Latin America and the Caribbean, the Middle East and North Africa, South Asia, and Sub-Saharan Africa. These regions comply with the “World Bank Country and Lending Groups” classification from The World Bank Group.[199]

Sentiment analysis

We used Valence Aware Dictionary for Sentiment Reasoning (VADER) to assess whether there was a positive or negative sentiment within a tweet.[94] The primary metric used for the sentiment analysis was the compound score (polarity), a unidimensional and normalized measure of sentiment between −1 and +1.

Topic extraction

We applied a K-means algorithm to the tweets in each region and gave each cluster a label according to the 20 closest tweets to the topic center and the most frequent words (top words) in the cluster.[108,200]

Emotion analysis

To determine the predominant emotion in each tweet, a classifier was developed based on texts focusing on four emotions: fear, anger, joy, and sadness.[201] We applied this classifier to all tweets to predict the probability of a tweet belonging to each of the four emotions.

Every Algorithm used for this study is available on Github: https://github.com/Chbour/Global_diabetes_burden. More details about the methodology can be found in Appendix 4.

Role of the funding source

The content of this publication is solely the author's responsibility and does not necessarily represent the official views of the funders.

Results

Spatial distribution of diabetes-related tweets

After preprocessing, we included 820,615 geolocated tweets in this study. Tweets were distributed as follows: 568,020 from North America (N=69.2%, 3 countries included), 176,124 from Europe and Central Asia (N=21.5%, 49 countries included), 31,426 from East Asia and Pacific (N=3.8%, 27 countries included), 20,465 from Sub Saharan Africa (N=2.5%, 36 countries included), 15,935 from South Asia (N=1.9%, 8 countries included), 4,554 from Latin America and the Caribbean (N=0.6%, 29 countries included) and 4,091 from the Middle East and North Africa (N=0.5%, 20 countries included). Figure 5.2 displays the distribution of tweets in each region.

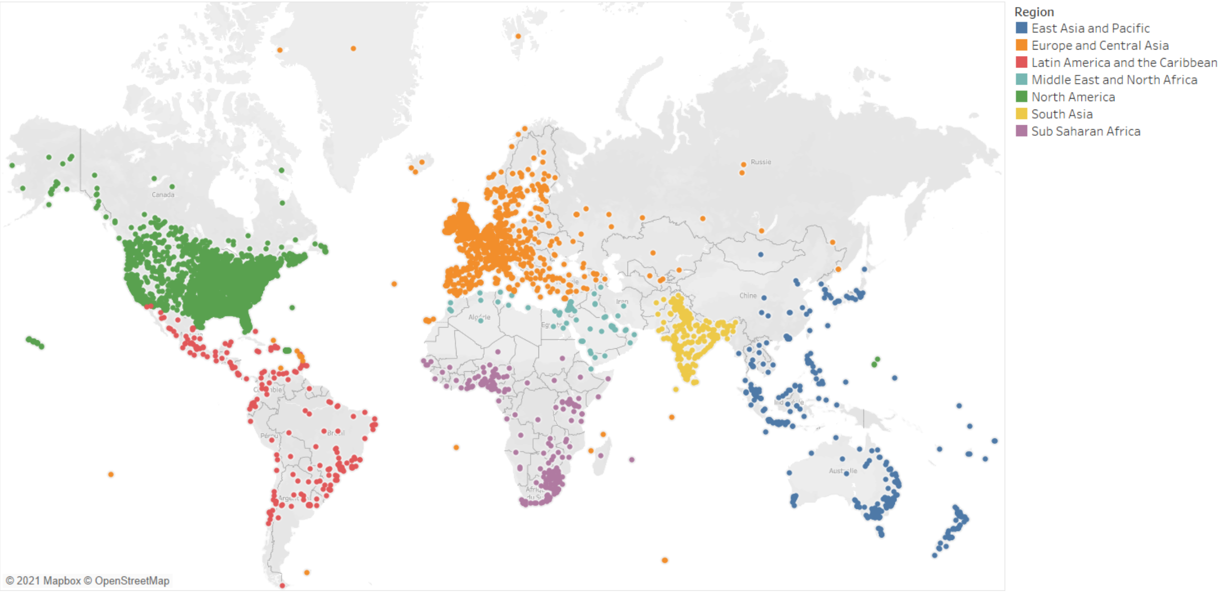

Region
🟦 East Asia and Pacific
🟧 Europe and Central Asia
🟥 Latin America and the Caribbean
🟦 Middle East and North Africa
🟩 North America
🟨 South Asia
🟪 Sub Saharan Africa
Canada
Russia
China
Algeria
Australia
© 2021 Mapbox © OpenStreetMap

Figure 5.2: Map showing the distribution of diabetes-related tweets according to the region (N=820,615)

Topics of interest

Among all tweets, 269,323 (32.8%) were predicted as posted by men, 311,343 (37.9%) by women, and 239,949 (29.2%) from unknown sex; 254,564 (31%) were from people with type 1 diabetes, 94,948 (11.6%) from type 2 diabetes and 471,203 (57.4%) from people where diabetes type was impossible to predict. Females were over-represented in East Asia and Pacific, Europe and Central Asia, Middle East and North Africa, and North America. Men were over-represented in Latin America and the Caribbean, and South Asia. In all regions, tweets identified as type 1 diabetes-related were predominant.

We identified four topics of interest for the people with diabetes from the Middle East and North Africa, 6 for South Asia, 8 of interest for East Asia and Pacific, 7 for Latin America and the Caribbean, 10 for Europe and Central Asia, 14 for Sub-Saharan Africa, and 18 for North America. They are further described below for each region and in Appendix 5.

"Glycemic Control" was a topic found in all regions. Six out of seven showed a common interest such as “Family and relatives” and “Food”, whereas “Insulin” matched for 5 regions. 4 regions had common topics related to “Comorbidities”. The significance of comparing percentages among emotions in topics in each region was determined using a Student t-test. All P-values shown are two-tailed.

Overall, South Asia had the most positive diabetes-related tweets and was associated with a higher polarity score while Latin America and the Caribbean had the most negative ones and were associated with a lower score (Table 5.1). Of the 820,615 included tweets, 356,683 were identified as positive (N=43.5%) and 308,811 were identified as negative (N=37.6%). South Asia and Europe and Central Asia had a higher proportion of positive tweets (resp. 47.6% and 46%). Latin America and the Caribbean, and North America had a higher proportion of negative tweets (resp. 38.2% and 38.5%). As shown in Table 5.1, the South Asia region was associated with a higher average polarity score, while Latin America and the Caribbean were associated with a lower score. The averaged sentiment scores were slightly positive and between 0.01887 (Latin America and the Caribbean) and 0.10376 (South Asia). Most regions had a positive score (greater than 0.05). In contrast, Latin America and the Caribbean, and North America had a neutral score (between -0.05 and 0.05) as these regions had a higher proportion of tweets with negative sentiment scores.

Region	**Mean sentiment score**	**Number of tweets with negative, neutral, and positive sentiment scores**
East Asia and Pacific	0.06961	Negative: 11,189 (N=35.6%) Neutral: 5,860 (N=18.6%) Positive: 14,377 (N=45.7%)
Europe and Central Asia	0.07209	Negative: 63,205 (N=35.9%) Neutral: 31,902 (N=18.1%) Positive: 81,017 (N=46%)
Latin America and the Caribbean	0.01887	Negative: 1,741 (N=38.2%) Neutral: 938 (N=20.6%) Positive: 1,875 (N=41.2%)

Middle East and North Africa	0.07022	Negative: 1,431 (N=35%) Neutral: 804 (N=19.6%) Positive: 1,856 (N=45.4%)
North America	0.02792	Negative: 218,717 (N=38.5%) Neutral: 108,246 (N=19.1%) Positive: 241,057 (N=42.4%)
South Asia	0.10376	Negative: 5,198 (N=32.6%) Neutral: 3,149 (N=19.8%) Positive: 7,588 (N=47.6%)
Sub Saharan Africa	0.05002	Negative: 7,330 (N=35.8%) Neutral: 4,182 (N=20.4%) Positive: 8,953 (N=43.7%)

Table 5.1. Average sentiment score and distribution sentiment scores. A sentiment score is considered negative, when lower or equal to -0.05, positive when greater than or equal to 0.05, and considered neutral when strictly between -0.05 and 0.05.[94]

East Asia and Pacific

On average, topics referring to users sharing support and advice such as “Type 1 diabetes communities” (48% compared to 31.4% on average in all other topics, *p<0.001*) and “Glycemic control” (39% compared to 31.6% on average in all other topics) were associated with higher rates of joy (*p<0.001)* but also with higher rates of fear (respectively 16.9% and 14.6% compared to 12.1% and 12.3% on average in all other topics, *p<0.001*) due to frequent fears about the future. “Insulin affordability” was associated with a higher rate of anger (28% compared to 16.6% on average in all other topics, *p<0.001*) because of users reacting to the huge insulin pricing gap between the United States and East Asia and Pacific.[202] “Diabetes-related complications and family history” was associated with a higher probability of sadness (45.8% compared to 38.1% on average in all other topics, *p<0.001*).

Europe and Central Asia

The two topics dealing with insulin (“Insulin access” and “Insulin and insulin supplies”) were associated with a higher probability of anger (respectively 28.6% and 26.2% compared to 15.87% and 16.3% on average in all other topics, *p<0.001*). Topics discussing relatives' life with diabetes and complications ("Diabetes-related complications and family history" and "Life changes since diagnosis") were associated with sadness (respectively 45.6% and 43% compared to 35.6% and 35.9% on average in all other topics, *p<0.001*). Topics "Daily management of diabetes" and "Type 1 diabetes communities" were mostly associated with joy (resp. 43.7% and 50.4% compared to 32% and 33% on average in all other topics, *p<0.001*).

Latin America and the Caribbean

Similar to Europe and Central Asia, the topic "Insulin issues" was associated with a higher probability of anger (28.7% compared to 15.6% on average in all other topics, *p<0.001*). Topics in which users shared love and advice ("Love and support" and "Glycemic control") were associated with a higher probability of joy (resp. 46.02% and 37.9% compared to 29% and 29.1% on average in all other topics, *p<0.001*). Finally, topics dealing with relatives' health complications and life with diabetes ("Complications and comorbidities" and "Experiences from relatives living with diabetes") were associated with a higher probability of sadness (resp. 47.9% and 47.8% compared to 42.7% and 40.4% on average in all other topics, *p<0.001*).

Middle East and North Africa

Topic "Insulin and insulin supplies" was associated with a higher probability of anger (28.8% compared to 15.6% on average in all other topics, *p<0.001)*. In this topic, users were reacting to the difficulty of insulin and insulin supplies self-management. However, sadness was the main identified emotion in all topics (39% on average).

North America

The five topics dealing with insulin pricing and affordability ("Inability to afford insulin", "Consequences of insulin unaffordability", "Insulin prices increase", "Insulin pricing including insurance" and "Costs implied by diabetes management") were associated with a higher probability of anger (between 20.1% and 32.7% compared to 17.9% to 18.8% on average in all other topics, *p<0.001*). Most topics were associated with a higher probability of sadness (41% on average) except "Type 1 diabetes communities", "Glucose tests", and "Sharing daily life" were associated with a higher probability of joy (resp. 46.1%, 48.7%, and 42.5% compared to 29.8%, 29.6% and 28.9% on average in all other topics, *p<0.001*).

South Asia

The highest average of anger was associated with the topic “Insulin use” (25.4% compared to 13.9% on average in all other topics, *p<0.001*). “Food habits” was associated with joy (39.01% compared to 30.2% on average in all other topics, *p<0.001*), while all other topics were mainly dominated by high rates of sadness (more than 40%).

Sub Saharan Africa

The topic “Insulin” was associated with anger (22.5% compared to 15.3% on average in all other topics, *p<0.001*) because of users’ angry reactions to diabetes misunderstanding and struggles to get insulin. The topic dealing with “Glucose guardian” was dominated by joy (39.3% compared to 28.9% on average in all other topics, *p<0.001*) as users were thanking others for their help or shared excellent glucose levels. In comparison, all other topics were dominated by sadness (between 37% and 46.02%).
Details about the average probabilities of sentiment distribution are available in Appendix 5.

Conclusion

In this study, we used worldwide social media data to better assess the global diabetes burden, from the perspective of PWD, and to study regional differences, which will serve to design more patient-centered diabetes programs. Social media data provide direct access to individual points of view and experiences of PWD, which can improve our understanding of how diabetes impacts their daily lives.

We have shown that some concerns are universal and shared by different online communities of PWD, while others are region-specific (e.g. North America, which has five insulin-related topics). We found that matters related to food, glycemic control, family and relatives, insulin, and comorbidities were shared by at least 4 of the 7 regions. Tweets in which users shared their concerns and experiences about their relatives’ diabetes, family health history, and comorbidities were associated with higher rates of sadness (47.2% of all related clusters and regions combined compared to 38.7% on average). On the contrary, most joyful tweets referred to users sharing advice, motivation, and peer-supporting and encouraging each other (37.7% of all related clusters and all regions combined compared to 31.1% on average). We also observed that 5 out of the 18 topics of interest in North America were related to insulin pricing, unaffordability, and the consequences of such pricing on health (on physical and mental health). Overall, these tweets correspond to 18.95% (N=101,019) of all tweets originating from the United States (N=532,981).[203] Additionally, these topics were associated with higher rates of anger (28.04% compared to 19.2% on average in the United States and 19.1% in North America). Meanwhile, users from Europe and Asia and other regions (Europe and Central Asia, East Asia and Pacific) were sympathetic to patients from the United States, sharing their disgust and misunderstanding of the insulin pricing gap between their region and the United States. These results from North America are consistent with the previous work from *Ahne et al.,* who showed that insulin pricing is a central concern among PWD on Twitter in the United States.[88]

Presumably, no previous study relied on such an extensive international database of posts from PWD to describe the diabetes burden. Our approach is more inclusive than those relying on questionnaires, such as Patient-Reported Outcome Measures or Patient Reported Experience Measures scales with predefined items. We monitored key diabetes-related concerns of PWD and quantified the associated emotions in different communities around the world. We have observed an elevated global burden of diabetes, with regional specificities that need to be taken into account more diligently.[204] Diabetes-related distress is present in every diabetes community and is sometimes under-researched, such as in Sub-Saharan Africa, and social media can help overcome these concerns.[205] *Özcan et al.* studied people with type 2 diabetes from different ethnicities in the Netherlands and showed that ethnicity is independently associated with high diabetes distress.[206] However, *Gariepy et al.* showed that diabetes distress in people with type 2 diabetes potentially varies according to some geographical and sociodemographic factors (such as social and physical order or cultural and social environment), which reinforces our hypothesis to compare diabetes burden determinants in different regions of the world.[207] Besides, patients' state of mind heavily influences their self-management habits. *Richman et al.* showed that positive emotions were associated with overall better health status, whereas *Coccaro et al.* suggested that diabetes distress is associated with negative emotions and the regulation of emotions.[208,209] Thus, as recommended by *Kalra et al.*, tackling patients' intellectual and emotional needs would be one solution to overcome the psychological barrier to adherence and self-care.[87] Our findings corroborate earlier research, indicating that diabetes burden is a common issue discussed on social media in all different regions of the world, and at different levels of severity. These findings also suggest that diabetes self-management is one of the biggest concerns, as PWD from the 7 World Bank Regions shared concerns regarding glycemic control and food. Moreover, concerns at the regional level were identified, such as insulin pricing in North America or the fear of complications and comorbidities in Latin America and the Caribbean. This discovery highlights the need to develop new global methodologies to tackle universal concerns regarding self-care and focus on more specific ones at a regional or country level to improve PWD experiences and deal with their outcomes.

This study has several limitations. First, the list of the diabetes-related keywords we used to collect the tweets may have been incomplete. This list has been created by translating an original list of English keywords, and we may have missed specific local diabetes-related keywords and associated issues in some countries. Second, some language-specific subtleties may have gone astray, as translating non-English tweets to English may obscure the original meaning. Third, although this study essays the diabetes burden on a global level, we did not manage to recover data from every country. However, this is the most comprehensive analysis on an international scale to date. Fourth, a bias in the geolocation analysis might exist, as the location is self-reported by users. We manually excluded areas that appeared to be fake. Furthermore, the geographical coordinates provided by a tweet's metadata were identified as being, by default, in the center of the country. As a result, the distribution map of the tweets shows geographical markers that are not necessarily in populated areas. Fifth, the precision of the different classifiers we used was not perfect. An additional limitation is that our results are based on subjective statements from people using social media and do not represent all PWD. Finally, due to the prevalence of sarcasm and irony on social media and the fact that we searched to define key emotions in every tweet, we cannot ensure that all emotions were correctly identified, despite our efforts to remove jokes and irony.

In this work, we demonstrated that the global needs and concerns of PWD varied vastly based on region and that the diabetes burden was perceived differently, despite some shared concerns. Our results suggest a necessity to improve the integration of these regional and global factors into future diabetes programs to enhance patient-centric diabetes research and care from the perspective of people with diabetes. This will contribute to improving the personalization of diabetes care and self-management.

Take-home messages from Chapter 5

1. A total of 820,615 geolocated tweets were included in the study. The distribution of tweets by region was the following:
    - North America: 568,020 tweets (69.2%)
    - Europe and Central Asia: 176,124 tweets (21.5%)
    - East Asia and Pacific: 31,426 tweets (3.8%)
    - Sub-Saharan Africa: 20,465 tweets (2.5%)
    - South Asia: 15,935 tweets (1.9%)
    - Latin America and the Caribbean: 4,554 tweets (0.6%)
    - Middle East and North Africa: 4,091 tweets (0.5%)
2. Each region had specific topics and emotions associated with diabetes-related tweets.
3. Common topics included glycemic control, family and relatives, and insulin.
4. Emotions varied across regions, with joy, sadness, and anger being prominent.
5. Social media data provided insights into the global diabetes burden from the perspective of PWD.
6. Regional differences were observed in concerns and emotions related to diabetes.
7. Topics such as insulin pricing and affordability were particularly prevalent in North America.
8. Diabetes-related distress and self-management were common concerns across regions.
9. Addressing both intellectual and emotional needs of PWD can help improve self-care and adherence.

Link with the following chapter

The analysis of diabetes-related tweets presented in this chapter provides valuable insights into the sentiments, topics, and regional distribution of public discussions surrounding diabetes on social media. Building upon this understanding, the subsequent chapter introduces a novel methodology to extend such analyses and replicate traditional cohort studies using social media data. This innovative approach involves the development of a Python package that leverages timelines to create virtual digital cohort studies. By seamlessly integrating social media data into cohort studies, this methodology opens up new possibilities for understanding health-related phenomena, including diabetes, in a digital context. Through the exploration of this methodology, researchers can gain a deeper understanding of the dynamics of diabetes-related discussions, uncover emerging trends, and identify potential interventions for improved diabetes management and care.

Chapter 6

ALTRUIST, a Python package to emulate a Virtual Digital Cohort Study on Twitter

Submitted to IEEE Transactions on Big Data

My contribution to this Chapter: *Conceptualisation. Data collection. Data labeling and classification. Data analysis. Data interpretation. Results interpretation. Package development. Drafting the article. Article revision.*

Abstract

Epidemiological cohort studies play a crucial role in identifying risk factors for various outcomes among participants. These studies are often time-consuming and costly due to recruitment and long-term follow-up. Social media data has emerged as a valuable complementary source for digital epidemiology and health research, as online communities of patients regularly share information about their illnesses. Unlike traditional clinical questionnaires, Twitter data offers unstructured but insightful information about patients' feelings and disease burden. Yet, there is limited guidance on analyzing Twitter data as a prospective cohort. We presented the concept of virtual digital cohort studies (VDCS) as an approach to replicate cohort studies using social media data. In this paper, we introduce ALTRUIST, an open-source Python package enabling standardized generation of VDCS on Twitter. ALTRUIST facilitates data collection, preprocessing, and analysis steps that mimic a traditional cohort study. We provide a practical use case focusing on diabetes to illustrate the methodology. By leveraging social media data, which offers large-scale and cost-effective information on users' health, we demonstrate the potential of VDCS as an essential tool for specific research questions. ALTRUIST is customizable and can be applied to various use cases, complementing traditional epidemiological methods and promoting minimally disruptive health research.

Keywords: Social Media, Digital Health, Python, Coding, Open Source, Diabetes, Software package, Algorithm, Machine Learning, Artificial Intelligence, Data Science, Computer Programming, Natural Language Processing, Text mining, Cohort

Introduction

This paper introduces a methodology and a Python package to generate and analyze cohort-like data using Twitter. A cohort is a study design that aims to conduct research in human populations and that helps to advance epidemiological knowledge. They are longitudinal studies in which research participants and numerous elements of their life (e.g. health, and social issues) are followed over time [14]. The methodology of such a study can be divided into five steps: 1) recruitment of the participants out of the target population, 2) obtention of baseline data on the exposure, 3) selection of the population, 4) follow-up and identification of the outcome of interest and 5) data analysis based on exposure and outcome of interest [210,211].

However, traditional cohort studies have several limitations. First, they are both time and cost-consuming as it can take years from the recruitment of the cohort population to the acquisition of sufficient data for analysis. Second, since the process can take several years, cohorts have a relatively long time of return on investment. Third, they are not suitable to monitor rare diseases or diseases with a long latency. Fourth, it might be difficult to maintain the participants' follow-up and limit attrition over time. [211,212]

The concept of a virtual digital cohort study (VDCS) was recently introduced by our team as a methodology complementary to traditional cohorts, based on social media data [213]. The large volume of data generated online by individuals becomes more and more relevant for the analysis of health-related topics such as recruitment or data analysis [43,128]. In particular, more than 500 million tweets were sent each day in March 2023 on Twitter, where a subset of public tweets can be accessed thanks to the Twitter Application Programming Interface (API) [89,214]. Twitter brings together communities of patients who share their experiences and feelings about living with a disease. By identifying these online communities, it is then possible to access and analyze a large number of spontaneous tweets describing the main concerns of patients [215]. Online data is also complementary to traditional data as it allows targeting specific populations that rarely occur in traditional data, such as minorities or people avoiding healthcare professionals [216,217]. Relying on existing online data from Twitter could circumvent some of the above-mentioned limitations (recruitment, cost, duration) while being compatible with traditional analysis methods, in particular survival analysis (such as but not restricted to Cox proportional-hazards models). Thus, combining both a virtual and a real-life cohort study, sequentially or simultaneously, could help conduct more relevant and patient-centric research.

Methods

ALTRUIST stands for virtu**AL** digi**T**al coho**R**t st**U**dy on Tw**I**tter u**S**ing py**T**hon. It is a Python package that aims to emulate a cohort on Twitter data. ALTRUIST was implemented and tested in Python 3.8 [218], which is an easy-to-use and open-source programming language that provides compact, readable, and portable code. ALTRUIST is open-source, available under GNU GPL v3 license on Github: https://github.com/Chbour/ALTRUIST. This package provides 1) scripts to collect Twitter data using the user's login information to the Twitter API, 2) separate personal and non-personal tweets to identify users with the outcome of interest, 3) collect specific users' timelines (i.e. history of all tweets of a user) and preprocess these timelines. 4) allows users to select the right Sentence-Transformer model for analysis (model that generates vector representations for sentences and can be used for multiple Natural Language Processing tasks [104]), 5) format the data to put them in a "cohort study" format, and 6) apply Cox proportional-hazards regression models (later in the text, simply “Cox models”) on these data. The main steps are described in Figure 6.1. This package is divided into several Python files. A Jupyter notebook file (.ipynb) is provided to illustrate how to call the functions and sequence the different steps. This allows the end user to easily link steps together and reproduce a VDCS on the topic of their choice.

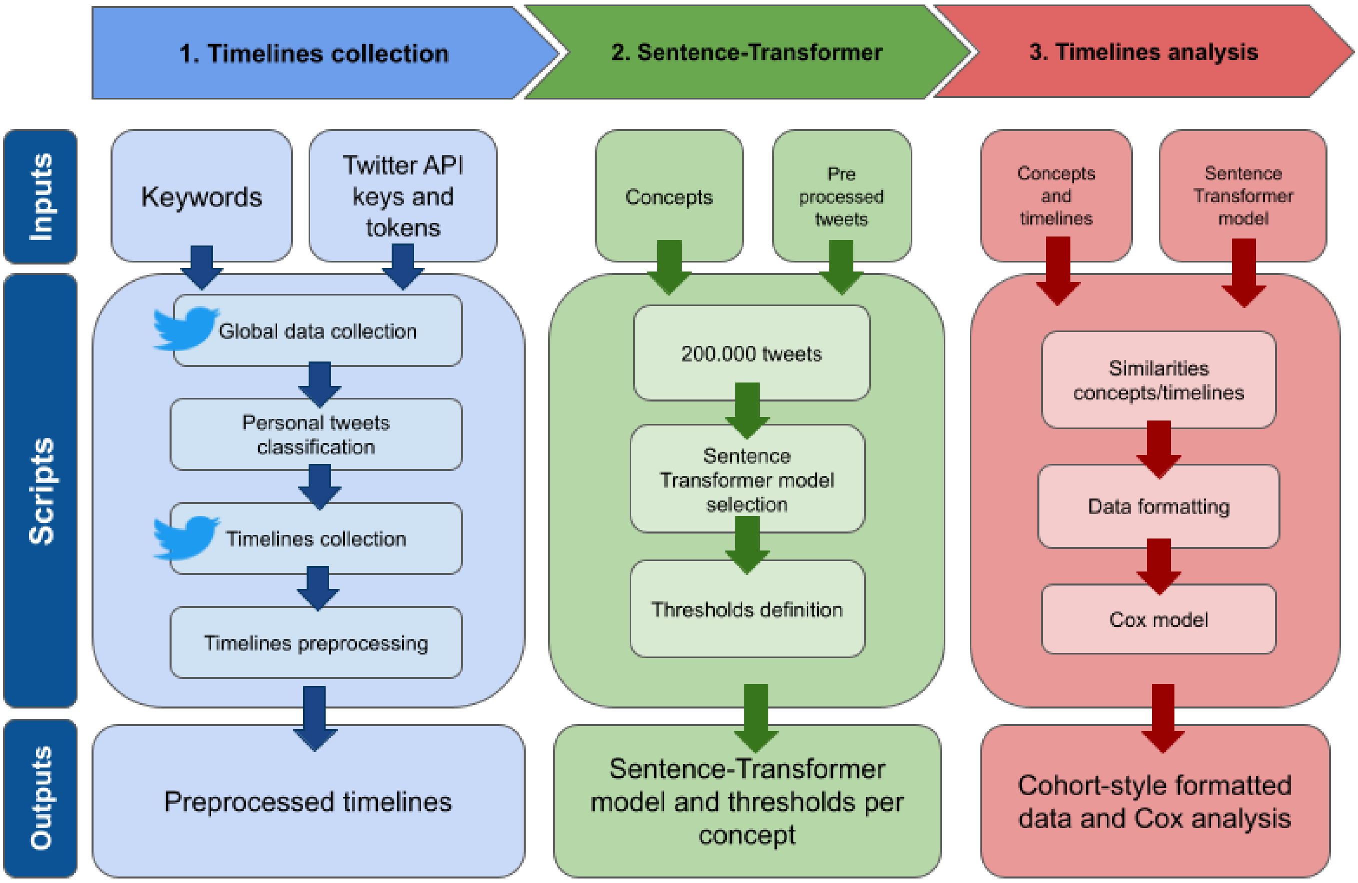


Figure 6.1. ALTRUIST structure and sequential workflow.

Data collection

Original data collection

The first step is to access the Twitter API [89] by making a request to Twitter explaining the use case and then to an application using Twitter API v1. Connection keys and tokens will then be generated, filled in the connection_data.txt file, and used to stream Twitter data. Tweets will be stored on MongoDB, a “powerful and scalable data storage” [91]. NoSQL-Database MongoDB, is particularly suited to store unstructured data such as text data from Twitter Connection data for MongoDB were also stored in the connection_data.txt file.

A list of keywords describing or related to the disease under study needs to be defined to collect tweets related to the topic, including relevant hashtags. The list of keywords need to be filled in the *keywords.txt* file. These keywords should be as exhaustive as possible in order to collect a maximum of relevant tweets. As the API is case-sensitive (distinguishing between uppercase and lowercase characters) the keywords should include words with lowercase and uppercase to be as exhaustive as possible. Any public tweet including at least one of those will be collected.

Population identification

The dataset thus created is, by design, mixing institutional content (tweets published by organizations, advertisements, research news, etc) and tweets with personal content. A tweet was considered personal if the user expressed his feelings or experiences about dealing with his own disease. As we aim to identify individuals tweeting about their own experience with the topic of interest, we need to create a classifier to identify personal tweets about the user's own experiences. To do so, some tweets have first to be manually labeled to identify and separate personal from institutional content. By default, we propose to fine-tune a *Bidirectional Encoder Representations from Transformers* (BERT) model, a machine-learning architecture for natural language processing pre-training developed by Google [99], and particularly its version *BERTweet*, a pre-trained model for English tweets [103]. For projects in other languages than English, a user can select another pretrained model, for instance from the Huggingface model hub which is pretrained on the specific language. The fine-tuned classifier is then applied to tweets in order to identify only the tweets containing personal information, meaning users tweeting about their personal experiences regarding the topic of interest are eventually identified. Note, in the next step, the entire history of all accessible data of this user on Twitter, namely their timelines, will be collected.

Timelines collection

Between the start of the original data collection and the beginning of the VDCS, users may have deleted their accounts or been suspended. The ALTRUIST package then automatically checks whether inactive accounts need to be removed before launching the timelines collection. Data from all remaining users can then be collected, excluding retweets. Retweets are not collected because they are not words directly expressed by the user.

Preprocessing

Once collected, the timelines are preprocessed in several steps. Several metadata fields are collected along with the tweet’s content. These fields provide additional information about the tweet, its author, and its context such as the language, the date of creation, and geographic location data. First, for each tweet, the *full-text* field in the tweet’s metadata is retrieved, which contains the full tweet text, and URLs and user mentions are deleted from it. Second, non-English tweets are translated into English using the package deep-translator [93]. Third, contractions are replaced (e.g. “can’t”: “cannot”). Fourth, dates are formatted to DateTime format (e.g. 'Mon Nov 23 13:52:51 +0000 2020' to “datetime.datetime(2020, 11, 23, 13, 52, 51, tzinfo=datetime.timezone.utc)” [219]). Fifth, empty tweets and those with less than 7 tokens were removed because we hypothesized that such short tweets do not contain any relevant information. Sixth, we define the acquisition date of each user and delete the tweets prior to this date. It is the first time a keyword is mentioned or the first time the cosine similarities between the tweets and the main concept exceed a specifically defined threshold. Cosine similarity measures the cosine of the angle between two non-zero vectors of an inner product space that measures. We use it to identify documents that are semantically similar to each other.

Preprocessed timelines from users with less than 100 tweets were deleted. Indeed, short timelines usually come from users who tend to share little information about themselves.

The following section describes the choice of the model to embed the tweets and compute the best cosine similarities between tweets and concepts, the definition of the threshold, and the calculation of the similarities for the timelines.

Sentence-Transformers

Vocabulary

Exposures (e.g: behavior, health state) and outcomes (e.g: death, comorbidities) need to be defined as concepts of interest. Each concept can be completed with several keywords to clarify and guide the research. For example, the concept "Mental health" can be completed by keywords related to depression or anxiety. It is up to the user to define these concept keywords and can be modified during the cohort if other exposures or outcomes of interest are raised.

Sentence Transformers algorithms

SentenceTransformers is a Python framework for state-of-the-art sentence, text, and image embeddings [104]. An embedding is a way of representing words or phrases as vectors of numbers, which can then be used as inputs to machine learning models. In this package, it is used to compute sentence/text embeddings and semantic textual similarity between tweets and concepts. *SentenceTransformers* provides a large set of pre-trained models including “All models”, which are general-purpose models that were trained on a large amount of available data. The package is automatically testing several of these models in order to allow the user to find the best-performing one. However, these models were not specifically trained on medical-related data. Thus, defining keywords associated with each concept will facilitate semantic textual similarities between tweets and concepts. To identify the model with the best performance, the cosine similarities between 200,000 preprocessed tweets and each keyword of the concepts must be computed using several models. The best model will then be used to compute textual similarities between tweets and the concepts.

Thresholds

For each concept, a threshold has to be determined manually by the user to decide when a tweet is related to a concept or not. This can be done, by using the previously chosen model and calculating the similarities between a large number of pre-processed tweets, here we chose 200,000 but more can be taken, all concepts/keywords. The similarity scores can be sorted in decreasing order for each concept, and a user can then screen the tweets with the highest similarities to identify the thresholds.

Timelines analysis

For each user, the beginning of the follow-up was defined as the beginning of the preprocessed timeline. The end of follow-up regarding a specific outcome was defined as the first date at which the threshold corresponding to the outcome was exceeded. If such a case does not appear, the end of the follow-up was chosen as the end of the timeline. Participation time is the number of days between the beginning and the end of follow-up.

Similarities between the preprocessed timelines from step one and the concepts defined in step two can then be applied to detect if users are tweeting about the concepts or not. Once these similarities are applied, we can move on to the timelines analysis. Each threshold crossing for the exposure between the beginning of the timeline and the last tweet was counted, for each quarter. We then convert this count to binary according to whether the user talks at least once about the exposure versus if the user never talks about it. The Python package “lifelines” is then used to implement a Cox model on the created dataset [220]. The output includes a hazard ratio and the associated p-value which can be interpreted. The hazard ratio is a measure used to compare how often the outcome happens in a subgroup of the population compared to another group [221].

Use case

This section provides a step-by-step use case to illustrate how ALTRUIST can be used to create a VDCS on diabetes. Our aim was to assess whether people with diabetes who talked about comorbidities had more mentioned frequent mental health issues. The *"Notebook_example_diabetes.ipynb"* notebook file illustrates the application of ALTRUIST and its functions to the use case of diabetes. This notebook can be directly modified by users to create new use cases.

To collect diabetes-related tweets, we defined 272 keywords in 28 different languages such as *diabetes, insulin,* and *blood glucose,* and related hashtags. These keywords were collected between 2017 and 2022 resulting in 34 million tweets containing at least one of these keywords. Details about the data collection, keywords, and the dataset can be found in our previous work [215].

Two authors (CB, GF) manually labeled 2,150 randomly chosen tweets into two categories: *personal tweets* (tweets containing personal information) and *institutional tweets* (tweets with non-personal information). We fine-tuned a BERT classifier model to keep only personal tweets written by people with diabetes.

The overall performances of our model on the test set are displayed in Table 6.1.

Accuracy	F1-Score	Precision	Recall
0.98	0.98	0.99	0.97

Table 6.1. Performances of the fine-tuned *BERTweet* personal content classifier on the test set.

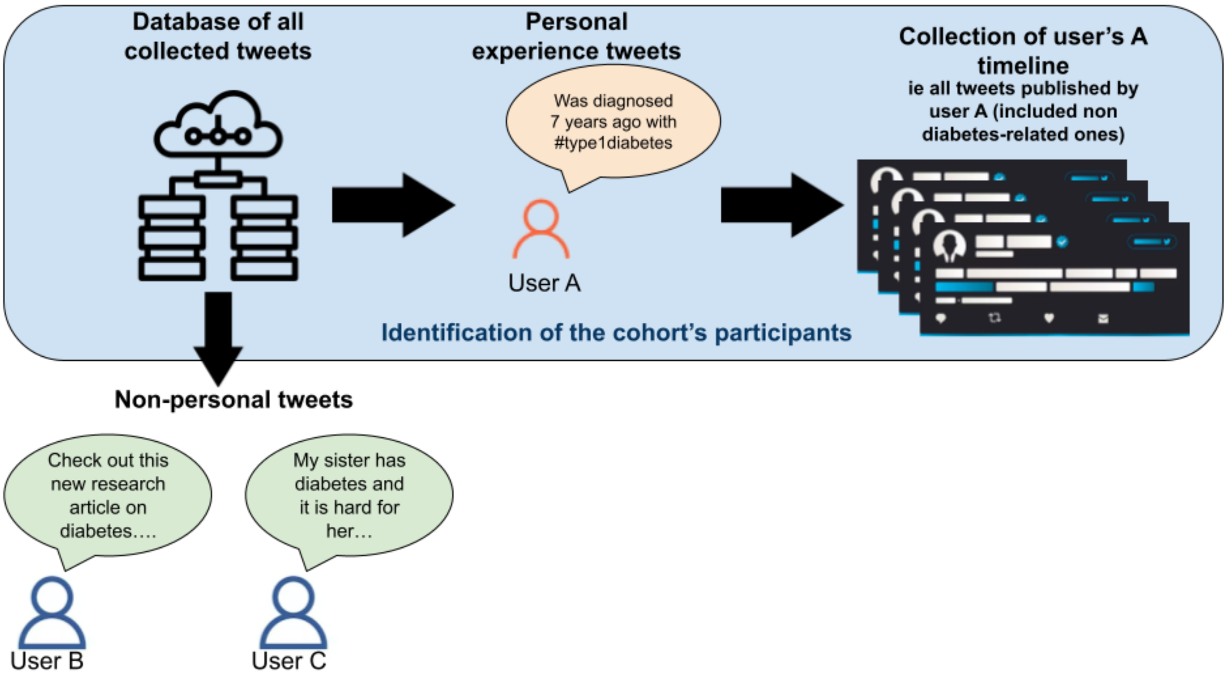

Database of all collected tweets
Personal experience tweets
Was diagnosed 7 years ago with #type1diabetes
User A
Collection of user's A timeline
ie all tweets published by user A (included non diabetes-related ones)
Identification of the cohort's participants
Non-personal tweets
Check out this new research article on diabetes....
My sister has diabetes and it is hard for her...
User B
User C

Figure 6.2. Identification of users tweeting about their personal experience with diabetes and collection of their timelines.

The 34 million diabetes-related tweets were then classified and users tweeting about their own diabetes could be identified. 88,057 users were identified after the classification step and 36,171 of them were still active at the time we launched our VDCS on diabetes. Approximately 60 million tweets were included for analysis after the collection and preprocessing of the timelines. Figure 6.2 shows the identification of the cohort's participants.

Based on previous studies [86,88,215], we defined a list of concepts related to the daily life, concerns, and health of people with diabetes. These concepts and associated keywords can be found in Appendix 6.

The next step was to identify the best model for our analysis. Table 6.2 shows an example of similarities between three short sentences and the key concept of "Diabetes" according to several models. A model was considered compatible with our use case if the similarities were different depending on the sentence. Indeed, some models computed almost identical similarities, whether the sentence was related to diabetes or not. The best model in our case was all-mpnet-base-v2. This model is built on the pretrained microsoft/mpnet-base model that has been fine-tuned on a 1B sentence dataset. This model was designed to be a sentence and short paragraph encoder [222]. We computed similarities between 200.000 preprocessed tweets and each concept using this model to define manually the thresholds. Thresholds can also be found in Appendix 6. We suggest testing this model first to check if it is suitable for further topics.

Model	Text	Similarities with concept “Diabetes”
roberta-base	Today, I’m going to try SentenceTransformers models to apply similarities between concepts and tweets.	0.914
	I was diagnosed with t1d two years ago. It has been really difficult to manage my insulin since.	0.908
	I struggle to keep my blood sugar levels stable all day…	0.929
all-mpnet-base-v2	Today, I’m going to try SentenceTransformers models to apply similarities between concepts and tweets.	0.116
	I was diagnosed with t1d two years ago. It has been really difficult to manage my insulin since.	0.397
	I struggle to keep my blood sugar levels stable all day…	0.417
bert-large-uncased	Today, I’m going to try SentenceTransformers models to apply similarities between concepts and tweets.	0.414
	I was diagnosed with t1d two years ago. It has been really difficult to manage my insulin since.	0.483
	I struggle to keep my blood sugar levels stable all day…	0.453
distilbert-base-uncased	Today, I’m going to try SentenceTransformers models to apply similarities between concepts and tweets.	0.408
	I was diagnosed with t1d two years ago. It has been really difficult to manage my insulin since.	0.443
	I struggle to keep my blood sugar levels stable all day…	0.408

Table 6.2. ***Examples of similarities between texts and the concept “Diabetes” according to several models.*** *For the Roberta-base model, the cosine similarities are always high, even when the tweet is not related to diabetes so we can exclude it. The bert-large-uncased and distilbert-case-uncased models always computed approximately the same values, so we can also exclude them. The all-mpnet-base-v2 model allowed us*

best to see differences in similarities depending on whether the sentence is related to diabetes or not.

Once the model and the thresholds were defined, similarities between timelines and concepts were applied. A cohort-style formatted table was then created and we applied the Cox model to it. For our analysis, we chose the concept of “Mental Health” as the outcome and “Comorbidities” as the exposure. The hazard ratio was HR=1.13 ($p<0.005$). In all, we have found that people with diabetes who report comorbidities are 13% more likely to report mental health issues.

Discussion

In this paper we presented ALTRUIST, a Python package to generate and analyze VDCS using Twitter data that can be used as a complementary approach to traditional cohort studies. It is open source and freely available on Github. Most of the tasks have been automated or semi-automated (i.e. requiring a decision from the user) which makes the use of this package and the sequence of steps easy to use. This package relies on Twitter data collection through the Twitter API and uses SentenceTransformers pre-trained models to compute semantic similarities with health concepts (exposure and outcomes). In particular, previous work by Klein et al. assessed the utility of Twitter data for a cohort study design. They concluded that Twitter can be a complementary resource for cohort studies to assess drug safety that can be analysed using LIWC [223]. VDCS that can be reproduced using ALTRUIST are free of use and not limited to a specific case. Indeed, ALTRUIST allows users to recreate all the different steps of a traditional cohort study on any outcome of interest. The package is easily tunable to the specific needs of each user and each use case. As the code is under an open-source GNU GPL v3 license, it can be modified and adapted. For example, the use of a BERT model is not mandatory and any other classifier can be trained and applied by users themselves. A package named Epicosm was recently introduced to link Twitter data with patients in existing cohorts [224]. It provides a way to collect timelines and analyze data using Language Inquiry and Word Count (LIWC) dictionaries which require a paid license [105]. ALTRUIST can be used as a complementary approach to this package once consents, user ids and potentially collect data to analyze the data.

Minimally disruptive clinical research is a principle that emphasizes the importance of developing epidemiological and clinical research studies with a focus on minimizing the burden on participants [225]. This principle suggests that certain epidemiological investigations should not be undertaken if we can already capture patient experiences and gain a better understanding of the impact of chronic diseases through online data studies. By using online platforms and collecting data remotely (such as social media data), researchers can gather valuable insights into the lived experiences of patients, without subjecting them to additional physical or psychological burdens associated with traditional research methods. This approach not only simplifies the research process but also ensures that the well-being of participants remains a top priority, while still yielding meaningful and comprehensive results. However, it is crucial to refer to the original methodological paper on VDCS to determine whether the research question can be adequately addressed using online data [213]. If so, the ALTRUIST framework can be employed. ALTRUIST can be used either as a standalone approach or in conjunction with traditional research methods, ensuring a comprehensive understanding of patients' experiences throughout the study.

Our use case, a VDCS on diabetes, apart from the data collection initiated by the World Diabetes Distress Study project [88], lasted less than two months which proves the efficiency and time-saving of this methodology. We found that people with diabetes who tweet about comorbidities are 13% more at risk to tweet about mental health issues such as depression and anxiety compared to those who do not. These results are consistent with what has been already published in the literature. Indeed, physical comorbidities (such as obesity and dyslipidemia) can have a negative impact on both mental health and quality of life [226–228]. Struijs et al. showed that people with diabetes with diabetes-related and non-diabetes-related comorbidities increase the health care demand [229]. We were able to show how people with diabetes-related comorbidities and who tweet about it are more at risk to mention mental health issues compared to people who do not talk about comorbidities. Depression is also a common non-diabetes-related comorbidity that increases the risk for diabetes-related complications [230]. These results suggest that ALTRUIST is a reliable tool to create and conduct VDCS on Twitter, that can complement traditional cohort study methodologies.

This package has several strengths. First, it is easy to configure and well-documented on GitHub. More than 30,000 participants were identified and included in our use case. We were able to collect data from these individuals and analyze them in less than 2 months, which would be difficult to do with a traditional cohort. Moreover, large and prospective population-based studies including at least 100,000 participants or more are called mega cohorts. For example, the UK Biobank study in the United Kingdom includes 500,000 participants [231], and the All of Us Research Program in the USA aims to reach up to 1 million participants [232]. Mega cohorts are long and expensive to set up, with a rather long return on investment. ALTRUIST also allows for some but not all use cases, of course, to reproduce mega cohorts in a time-saving way. Indeed, the more exhaustive the collection of tweets (i.e. complete keyword list, sufficiently long data collection), the more it will be possible to identify a large number of users talking about their own disease. Finally, the best model selection is based on the testing of several existing pre-trained models. It is possible for the user to add new models to be tested which keeps the methodology up-to-date and scalable.

This package also has several limitations. First, the analysis performed during the VDCS are based on subjective statements from people using social media and do not represent all people living with the outcome. Second, the larger the cohort population, the more time-consuming the preprocessing and similarities computation between timelines and concepts will be. Moreover, using transformer-based models is extremely heavy; using a GPU might be necessary to perform some of the tasks efficiently. Still, these processes, combined with the data collection and analysis, will take less time than any real-life recruitment would have taken in a traditional cohort setting. Third, the cosine similarities and the definition of the threshold per concept are also based on a subjective choice of the user. This step is difficult to automate because it is use-case specific. Finally, we tried to define the period of exposure and the date of acquisition as would have been done in a traditional cohort. However, the notion of duration is more complex with social media data because it depends on tweets and not on questionnaires that would be administered at predefined times. As such, the duration is therefore not the same for everyone and must be interpreted cautiously.

Conclusion

We developed an open-source Python package that facilitates the generation and analysis of VDCS on Twitter. It is an easy-to-use tool to add to the arsenal of health researchers to run digital epidemiology projects.

Take-home messages from Chapter 6

1. ALTRUIST is a Python package designed to generate and analyze VDCS using timelines-based social media data such as Twitter. It serves as a complementary approach to traditional cohort studies and is available as an open-source package on GitHub. Most tasks are automated or semi-automated, making it user-friendly and easily adaptable to different research needs.
2. ALTRUIST uses timelines-based social media data and SentenceTransformers pre-trained models for computing semantic similarities with health concepts.
3. VDCS created with ALTRUIST are versatile, allowing users to recreate various steps of traditional cohort studies for any outcome of interest.
4. Minimally disruptive clinical research principles emphasize the value of online data studies to understand patient experiences while minimizing burdens on participants.
5. ALTRUIST can be employed either as a standalone approach or in combination with traditional research methods to gain comprehensive insights into patient experiences.
6. The efficiency and time-saving nature of ALTRUIST were demonstrated in a VDCS on diabetes, yielding consistent results with existing literature.
7. ALTRUIST provides a valuable tool for health researchers to conduct digital epidemiology projects, adding to the arsenal of available methodologies.

Chapter 7

Discussion, conclusion and perspectives

As we increasingly embrace digital technologies in our daily lives, they are influencing our comprehension of various health issues, driving a paradigm shift in health research and care. This digital health research project explored the role and potential of social media for health research with a specific focus on diabetes.

This chapter will discuss the general findings of this work and their implications, critically assess the strengths and limitations, and highlight potential future directions in digital epidemiology.

General findings

The use of social media for health research purposes

The first objective provided a comprehensive examination of the use of social media for health research, showing a significant growth and diversification in the use of social media for health research over the past decade. Researchers harnessed the power of social media like Twitter, Facebook and Instagram for different health-related purposes. Social media were mostly used in research related to communicable and chronic diseases, mainly because these platforms are widely used and have a lot of users. The prevalent use of social media among young adults and adolescents provides a unique way to learn about the attitudes and health of these age groups, which could be used to improve their health in the future [233]. Moreover, the spontaneous nature of social media facilitates real-time monitoring, providing insights into evolving health-related events.

The uses of social media in health research also evolved. They moved beyond traditional use in studying diseases. Indeed, it highlighted the immense potential of social media to transform traditional research methodologies. The limitations of traditional recruitment strategies have been mitigated using social media, which offer a cost-effective and far-reaching alternative [234,235]. Facebook can be used as a tool for recruitment, as it offers researchers a unique opportunity to dig into a diverse population.

The exponential growth of big data, the advancements in machine learning and data mining techniques have also paved the way for health research using social media data. The profusion of generated data has proven to be valuable for disease surveillance, epidemiology and forecasting [196,236]. Particularly, the API provided by Twitter until 2022 allowed real-time data access and was one of the most effective tools for health research.

However, there is a need for ethical considerations and limitations that researchers must navigate while using social media. Ensuring user privacy and data security remains a challenge, given the public nature of social media data. Moreover, the potential biases due to self-selection, exclusion of certain demographic groups and potential multiple entries from the same user can be obstacles to research validity. For this field to fully realize its potential, we need to address the critical need for more guidance according to the different applications of social media for health research, both from a methodological and an ethical perspective. It is crucial to evaluate and refine these methodologies through the years as social media are rapidly evolving to ensure they are effective, inclusive and ethical.

Traditional questionnaires and surveys to monitor diabetes distress are both time-consuming and costly. Collecting and analyzing experiences from people with diabetes around the world would have required massive investments, logistical coordination and time. Social media hold high potential for enhancing our understanding of people with diabetes health and well-being. It also becomes particularly relevant in the context of minimally disruptive clinical research, since we can collect large amounts of data with minimal disruption to participants, use it as a complementary approach to traditional epidemiology and as such reduce the volume of data collected from the participants or self-reported by them. Indeed, rather than asking individuals to fill out extensive questionnaires or participate in lengthy interviews, we can derive insights from their experiences shared online, ensuring a patient-centric and efficient approach. Beyond the insights we obtained that are relevant per se, we showed that the methodologies developed under Objectives 2 and 3 can complement and sometimes reduce the research burden.

Building upon this potential, a second objective was developed to explore regional differences in diabetes burden as perceived by individuals living with the condition, using artificial intelligence approaches.

Twitter data to better understand diabetes distress

This research can be seen as a first step towards an online diabetes observatory, where millions of worldwide diabetes-related tweets were collected, processed and analyzed. This method allowed us to capture the real-life experiences, concerns and emotions of people with diabetes across the globe, providing an innovative patient-centered perspective compared to traditional questionnaire-based approaches [237].

Some key concerns of people with diabetes were found to be universal while others were region-specific. Issues related to food, glycemic control, family, insulin and comorbidities were common across the globe. This is consistent with previous studies that have shown that people with diabetes often struggle with dietary management and maintaining glycemic control. However, some concerns were region-specific. A first example are the topics related to insulin pricing in North America that Ahne et al. already showed to be one of the main concerns in the United States but also generally known [88,238]. In Europe and Central Asia, where healthcare systems often cover a part of medication costs, one of the main concerns was the daily management of diabetes. This supports the idea that self-management of diabetes is an important part of diabetes care [239]. In all, we need to consider regional contexts when designing health interventions or policies regarding diabetes.

Diabetes-related tweets also showed high levels of emotional distress linked to diabetes. Tweets related to diabetes complications, family health history, and comorbidities were associated with higher rates of sadness. This aligns with a previous study concluding that diabetes distress management should include a psychological approach with diabetes-specific care [240]. Finally, social media already proved to be used by people with diabetes for peer-to-peer support [241] and we found that tweets from these communities were associated with more positive emotions such as joy.

By harnessing the data generated by online communities of patients, valuable insights can be gained into the emotional burden experienced by individuals living with diabetes. Expanding upon this, the third objective aimed to implementate a new approach known as the VDCS, using an open-source Python package called ALTRUIST.

Virtual Digital Cohorts Studies using ALTRUIST

This new methodology leverages timelines-based social media data to conduct large-scale and mega cohort studies. The ALTRUIST framework provides an efficient, cost-effective and minimally disruptive methodology to emulate cohort studies using social media compared to traditional prospective cohort studies [242].

ALTRUIST was created to supplement existing methodologies in digital epidemiology, which often rely on predefined questionnaires or specific prompts to collect experiences from patients. The package uses social media such as Twitter, which do not rely on predefined questions. Thus, it provides a direct access to the unstructured, real-world experiences and concerns, and emotions of patients all around the world. ALTRUIST makes this process more straightforward and easy to use, marking a significant step forward in research methods. The package is openly available on Github for the community. Its accessibility ensures that researchers across various domains can benefit from its functionalities, promoting structured and reproducible social media data analyses. This tool makes social media data analysis more accessible and speeds up the analysis from large social media datasets.

Using this package on a use case on diabetes highlighted the potential of this approach. By analyzing timelines of people with diabetes, the study was able to identify how people with comorbidities were more likely to disclose mental health issues. This aligns with the existing literature on the fact that people with diabetes and comorbidities are presenting a worse mental health and patient reported outcomes [243–245]. It also shows the need to integrate social media data to gain a more nuanced and patient-centered perspective into research.

Perspectives

AI: a rapidly evolving field

AI is a rapidly evolving field. New AI techniques are continuously developed and could potentially be used to revolutionize how we analyze data. In the face of such rapid innovations, researchers must constantly adapt and revisit their approach to analyze the data to address their research questions. They need to stay up-to-date with these changes, so they're ready to handle new types of data and meet new challenges. Just 15 years ago, researchers started using social media for health research purposes. At first, basic ML techniques (e.g, linear regression, support vector machines, random forest) were used to analyze social media data. Such methods were efficient to spot patterns and trends but struggled to understand the messy and unstructured data from social media.

Then researchers started using word embeddings to convert words into vectors that machines can understand. This helped to better understand the meaning and context of words in health-related social media data. But even with word embeddings, there are still limitations.

An illustration of this rapid evolution in AI was also seen during the thesis itself. Three years ago, it was difficult to deal with multilingual datasets containing more than 2 languages. Many machine learning models were trained on English data due to the accessibility and abundance of such datasets. When researchers had multilingual datasets, a common approach was to translate all the data into English to leverage these pre-trained models, especially if they didn't have the resources to train a new model for each language. This was often seen as a practical solution, but also imperfect as nuances and context can get lost in translation. Moreover, freely available translation APIs often have request limits, making it a challenge to translate large datasets.Since then, the landscape of AI has radically changed. Huggingface developed the BLOOM model, the first multilingual Large Language Model (LLM) that can generate text in 46 languages. This means that we can now directly apply such models to multilingual datasets, such as our diabetes-related tweets, without the need for translation. For instance, while BLOOM is primarily an autoregressive language model designed to generate and continue texts, it is possible to use it for data clustering. Indeed, we could extract the embeddings for each tweet using BLOOM and then apply these embeddings to clustering algorithms, as we did in objective 1

The world of AI also continued to innovate and evolve, giving birth to even more advanced techniques in a few years. Transformers models, such as BLOOM, have proven to be highly effective to understand human language and to pick up the context of words in a way that was impossible before. The emergence of LLMs such as GPT-4 have opened up new ways of understanding and generating human-like text. In health research, this can help to better understand the feelings of patients, identify health-related misinformation and better understand public awareness on health conditions to orient public health campaigns. Since GPT-4 can be used to read medical records, analyze symptoms, and suggest diagnoses [246], it could also be used to analyze patients' social media timelines to better understand their health status and evolution. This can then help healthcare providers make more informed decisions and provide better patient care. Since LLMs can analyze large amounts of data, they may be used for disease surveillance and epidemiological analysis to identify trends, patterns, correlations, and risk factors in social media data [247]. For instance, a recent study showed that GPT models can be used to identify misinformation regarding COVID-19 vaccination [248]. Moreover, chatbots can also be created to directly exchange with patients online and support them in their daily life with a chronic condition [249]. Thus, large amounts of data can be analyzed using such new AI models to improve health research to a more personalized, informed and effective patient care.

Beyond Text: Visuals and Voices on Social Media

It is not just AI that is evolving: social media are also rapidly transforming. In the early days of 2005, blogs were the primary source of social data. Soon after, Facebook and Twitter have emerged, providing a lot of textual data to analyze for researchers. However, in recent years the focus has shifted towards more visual and less textual data: pictures and videos. Platforms like TikTok, Snapchat and Instagram have become increasingly popular, particularly among the youngest. However, these types of data are more complex and challenging to analyze. Indeed, it requires more computational power and more advanced analysis techniques. Current methodologies to analyze visual content from social media for health research are still at an early stage, and much research is needed to catch up with the rapid shift towards visual content. First, dealing with pictures and videos requires new ethical considerations and guidelines since pictures can contain identifiable features. Second, there is a need to collaborate between experts to process and interpret visual content as image processing and analysis has its own rules and guidelines. Third, not only the sole content of images can also be analyzed. In a previous study, Reece et al. showed that features extracted from Instagram photos can be used as predictive markers of depression regarding the colors and filters used [250].

Furthermore, the current trend in AI is towards voice. Platforms such as virtual assistants (e.g. Siri), smart speakers (e.g. Alexa) and smartphones are increasingly voice-based and can be considered as sources of vocal data. Such data is particularly interesting due to the potential of vocal biomarkers. These can be considered as indicators in the voice that might indicate health-related conditions or emotional states. AI for voice analysis is an opportunity for healthcare. It can be used to identify vocal biomarkers for risk prediction, and remote monitoring of various clinical outcomes and symptoms [251]. Moreover, just like textual data, collecting voice recordings supports minimally disruptive clinical research. It is a less invasive alternative to traditional sample collection in research studies. Still, working with voice data can be challenging. The voice is directly linked to an individual's identity, and is therefore considered sensitive data. There is also a need for clinical validation to ensure that the biomarkers are working and relevant for the clinical research [252]. In all, voice data in AI and health research could transform our approaches to diagnose and manage diseases.

Coming back to social media, platforms such as YouTube, Twitter, TikTok, and Instagram have a wealth of videos where individuals narrate and share personal experiences. These are called “storytimes”. Many of these narratives include discussions about their health conditions (e.g. depression, cancer, diabetes). Extracting audio recordings from such videos could provide a valuable source of voice data which could be used to identify potential vocal biomarkers, especially when individuals detail their symptoms or emotional experiences related to their conditions. Such an approach not only taps into the vast amount of user-generated content online but can also highlight real-world and spontaneous expressions of health-related issues. However, such voice data can be more complex to process. Unlike the standardized conditions of clinical settings, these recordings are varied and can introduce more nuances due to background noises, accents, and spontaneous speech patterns. With advanced AI and sound processing techniques, this data could pave the way for novel discoveries in health research using social media.

The dynamic of digital landscape

The social media that researchers rely on for data are not static. They evolve over time, providing new features both for users and researchers. A perfect example for this is

Twitter. Researchers have relied on the API for years to collect data for their studies. However in 2023, Twitter made significant changes to this API, changing the access rights for researchers and adding pricing to access the data. This could hurt the reproducibility of previous work conducted using the API and put an early end to searches using this social media platform. Another example is the Twitter update that allows paying users to publish longer tweets. This means that researchers who have been studying short text data may now have to adapt their methodologies and consider integrating large texts into their analysis. But this is also an opportunity. With extended text data, researchers have the potential to perform more nuanced analysis, capturing more details from the patients' perspective. On Reddit for instance, communities of patients are gathering in specific health-related groups called subreddits. In one of our studies called DOVA-Lux (Digital Observatory to monitor VAccination hesitancy in Luxembourg), we analyzed Long COVID related posts on specific subreddits (e.g. r/LongCovid, r/covidlonghaulers) to study the profiles of Long COVID from the perspective of the patients spontaneously sharing their experiences and symptoms on Reddit. By adapting the methodologies presented above to the processing of longer texts, we were able to characterize the heterogeneity of Long COVID profiles. General symptoms, such as fatigue, were the most prevalent and frequently co-occurred with other symptoms [253]. Moreover, Reddit also showed to be a great source of data to study depression and anxiety [254], to analyze topics and sentiment around lifestyle habits (such as vaping [255]) and to study chronic diseases [256].

These examples show that social media have their own business model and technical changes. Even if research is mainly focused on Twitter, other social media such as Reddit and Facebook provide API that can be used for data collection. Researchers thus need to be ready to adapt to the changes, and need to develop strategies that can accommodate the dynamic of the evolving digital landscape.

Integrating traditional methods to modern tools

Digital epidemiology has been an important shift in how we consider health research, offering a novel view to track, predict and understand health issues. However, while the potential of digital epidemiology is vast, it is still a relatively new field and has to keep growing to fully leverage the traditional methods of epidemiology. Social media proved to be a great tool for digital epidemiology, using real-time data analysis for surveillance, forecasting, to track health trends and predict outbreaks. However, epidemiology is not just about watching trends and making predictions. It is a key part of public health and is used to study how and why illnesses happen in a population. Epidemiologists help create plans to prevent sickness and improve population health. Thus, to fully tackle the power of digital tools, AI and ML models need to be refined and expanded to handle more complex and epidemiological-oriented analysis, using both digitosome and traditional omics. This includes studying the determinants of diseases or health, the interventions for disease preventions and studying the cause-effect relationships between outcomes and exposures. Another challenge is data integration as we need to combine different data sources such as social media data with questionnaires, medical record data or data from connected devices. By combining these datasets, we can have a clearer view of the overall health of patients and move towards the concept of precision health. It will not only provide a richer context but also more targeted health interventions tailored to individual needs.

While digital epidemiology has made a significant advancement in transforming health research, we are just at its beginning. The field must continue to evolve and grow, combining all the data sources available while assimilating more features from traditional epidemiology to reach its full potential.

Bridging the gap between digital epidemiology and healthcare practice

The overall goal is not to replace traditional epidemiology but to complete it by integrating the strengths of traditional methods with digital tools and data. AI is still often perceived as something enigmatic that will try to replace human taskforces, and this sentiment could also be shared in the healthcare sector. Some healthcare providers may not fully understand what digital epidemiology is, why we are doing it and why funds and resources are invested in this area. They may not see how it could potentially be used to improve patient care, inform public health decision makers and improve health outcomes. This lack of understanding may cause them to ignore the results generated through digital epidemiology studies. A solution could be to highlight how digital epidemiology directly impacts their work and their patients' outcomes. For instance, with the World Diabetes Distress Study, we shed light on diabetes distress, and were thus able to explain to diabetologists and general practitioners the extent to which it impacts on the lives of people with diabetes. Most of them were perhaps unfamiliar with the term, and unaware of the magnitude of the impact of diabetes-related distress. Thus, these social network analyses of millions of people with diabetes have helped to increase awareness of the subject.

In all, we need to work all together to improve the standards of healthcare by merging the unique human expertise of healthcare providers and the powerful data-driven approaches from digital epidemiology.

Limitations of social media data

Data protection aspects

While social media data holds high potential for health research, it is important to consider its limitations regarding privacy and consent. The General Data Protection Regulation (GDPR) is a data protection law in the European Union. It imposes requirements and guidelines on how to handle personal data. It includes lawful processing, data minimization and the right to erasure and thus should be considered when using social media data in research.

Consent is a major issue. Users share their health information and feelings on social media without knowing that it could be used for health research. On Twitter, the tweets are considered public data. This can lead to researchers thinking that every platform is the same and create a breach of informed consent. Thus, obtaining consent for the use of social media data when necessary can be important.

User’s privacy could also be at risk. It is sometimes possible to identify individuals even if the data has previously been anonymized. For instance, even if a tweet has no user names or mentions in it, writing the text in a search engine can easily redirect to its original author. This is particularly important when dealing with health-related data which can be considered as sensitive. To address such concerns, we established a set of best practices to ensure data privacy and ethics. We consulted the Luxembourg Agency for Research Integrity to validate our data analysis practices. Indeed research on social media can be a complex undertaking due to the fact that the data is often viewed as “public domain” but in fact, there are privacy and other issues that overlap. We thus created a set of good practices to guide our research. These practices suggest that 1) we do not own the data, we only obtained the right to access and use them, 2) we publish only aggregated data and do not share any individual post and 3) we do not exchange with people with diabetes directly on social media. We have also adapted and extended the good epidemiological practices, to move closer to a more rigorous, transparent, and ethical research. Another key challenge is the data access and re-uses. Data needs to be stored and protected as efficiently as possible and analysis needs to stick to the context and goal of the research project and not spread all over.

In all, we need to ensure that using social media for health research has to be realized responsibly and ethically.

Lack of standardized methodologies

Another challenge in digital epidemiology is the lack of standardized methodologies and reproducibility, whether it is for surveillance, forecasting or data analysis. In traditional epidemiology, well-defined and reliable methods were created to generate valid and reproducible results. However, in the world of digital epidemiology, each study has different approaches, making it difficult to compare or combine results. This is probably due to the wide range of data sources, from social media to digital health tools such as wearable devices. Each source has some specificities and challenges that require special methodologies for data collection and analysis, and cannot necessarily be freely shared. And while the field can be seen as innovative, it also needs to have standardized and regulated approaches. This will not only facilitate collaboration and comparison across studies, but also help to increase the impact of digital epidemiology in the healthcare providers and medical communities. Finally, as digital epidemiology keeps growing, it will become more difficult to develop these standard methodologies and have them adopted by researchers in the field, just like in traditional epidemiology. For instance, the Innovative Medicines Initiative WEB-RADR project, funded by the European Union, explored the value of social media in pharmacovigilance [257]. They provided several recommendations such as establishing a framework (to ensure clarity and patient safety), following the existing guidelines regarding the use of social media for health research, review the pros and cons of social media for pharmacovigilance and respecting patients privacy. Such projects are therefore a first step towards more standardized, transparent and reproducible methodologies.

In line with the goal of transparency and reproducibility, all the scripts used in this thesis have been made open-source and are available on GitHub, ensuring full clarity on the data processing and facilitating potential replication of the study by other researchers.

Next steps of this thesis

Worldwide Online Health Observatory

Between 2020 and 2023, billions of health-related data was gathered on Twitter. This includes tweets related to diseases (e.g. cancer, Alzheimer), medications (e.g. Proair, Ventolin), mental health (e.g. anxiety, depression) and symptoms (e.g. headache, cough) and provides an insight into global health-related discussions happening online.

The next step is to extend the analysis on diabetes burden to this global data collection of hundreds of millions of tweets to create a Worldwide Online Health observatory. This aims to provide a global overview of health-related trends and concerns around the world in almost real-time. It can be used to detect emerging health-problems, potentially enabling faster response in the case of health crises. Essentially, this observatory could be used as a live tracker of worldwide health trends, harnessing the power of social media data to monitor global health.

A preliminary proof of concept is currently under development to bring this observatory to life. This includes analyzing the health-related tweets, storing the aggregated results and loading them into Tableau to visualize these results. This will allow users to explore health-related trends and discussions from the previous week, providing a user-friendly way to understand the data. Figure 7.1 shows the proof of concept already developed.

With the recent pricing of Twitter's API, the data collection needed for such a project may no longer be feasible for academic institutions. Still, such data can be interesting for entities such as governmental organizations and health authorities. The insights derived from this platform could then be used to guide public health policies and interventions, thereby maximizing their impact on population health. Thus, despite the cost, the potential benefits could make it a worthwhile investment for such organizations.

This version provides several features. It includes a distribution of the health-related tweets around the world and a cluster analysis that groups tweets according to the identified topics. The observatory also provides the average sentiment analysis associated with each topic. A word cloud visualization is also available (for the overall collection or the individual clusters), offering a view of the most frequently used terms in the health-related discourse. The proof of concept thus shows the potential of this tool in exploring and understanding health-related trends and sentiments around the world in almost real-time.

Further work is needed to transform this proof of concept into an operational observatory. The first step includes automating the analysis and conducting them weekly. Then the aggregated results need to be stored in a secured server where the data can be accessed as needed. The second step involved making the Tableau-based observatory accessible to online users, allowing anyone to explore the visualizations and gain insights from the data. The third step involves expanding the observatory to other social media than Twitter, that do not necessarily require paid access. Incorporating other sources of data could provide a more comprehensive picture of global health trends and discussion. By integrating all these features and improvements, a comprehensive, reliable and easy to use Worldwide Online Health Observatory can be established. It will be a useful tool for public health agencies, researchers and policy makers to help them make informed decisions and in their efforts to improve global health.

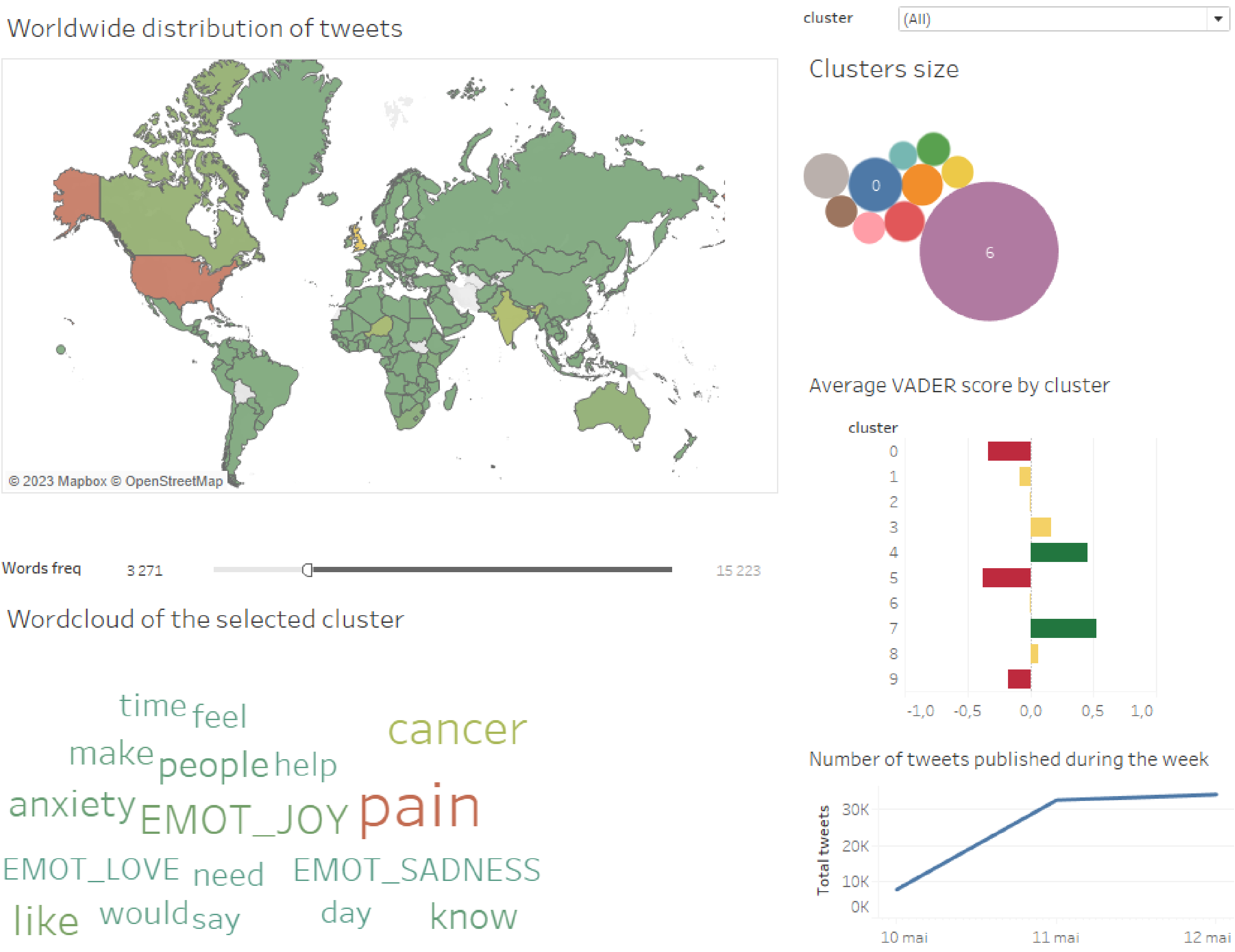


Figure 7.1. First proof of concept of the Worldwide Online Digital Health Observatory

Future versions of the ALTRUIST Package

Future updates of the ALTRUIST package will introduce new features to take care of the evolving needs of its users. One update will enable users to select the social media from which they want to collect data, thus broadening the sources of data to Reddit for instance. Another update could integrate more methods to compute similarities between text and concepts, along with the advancements in the NLP models. Potential updates could also add the ability to handle larger datasets more efficiently. But since it's not just user needs that are evolving, but also AI methods, it would be interesting to integrate new models such as multilingual models or GPT-4 into the package. Integrating multilingual models would make it possible to create multilingual cohorts and to focus on certain populations. GPT-4 could be used for language translation (if needed) and for natural language understanding, such as computing the similarities. Each of these enhancements would make the package more versatile and robust, offering a comprehensive tool for digital epidemiologists.

Conclusion

We have shown that using social media data allows us to explore the daily sentiments and concerns of people with health conditions like diabetes. It is a treasure trove of information for researchers to gain insights into the patients experiences and outcomes. By harnessing the power of social media using advanced analytical methods such as AI and NLP, researchers can break free from the limitations of traditional research methods.

However, this remarkable opportunity must be accompanied by a sense of responsibility and the desire to follow guidelines. Just as research in traditional settings follows strict protocols, there is a need for a more ethical and responsible framework for research using online data. Adhering to these guidelines will assure the preservation of privacy, facilitate the application of proper informed consent procedures, and rule careful data management practices. This way, we can maintain the rigor, reproducibility and trustworthiness of our research while simultaneously protecting the rights and interests of the individuals involved. Moreover, integrating digital data into research requires collaborative work between researchers and healthcare providers to tackle the challenges of digital data and health research at the same time, to have a real impact on patients.

In the end, the use of digital data such as social media data in health research is a game-changer. It pushes us into new areas, uncovering unique aspects of human experiences and setting the stage for future health improvements. By accepting this change and taking control of these opportunities we can create new methodologies, ethical rules and guidelines that can transform healthcare and make a future where everyone's health gets better thanks to their own digital data.

References

1. Principles of epidemiology. 20 Dec 2021 [cited 2 Mar 2023]. Available: https://www.cdc.gov/csels/dsepd/ss1978/lesson1/section1.html

2. [No title]. [cited 2 Mar 2023]. Available: https://dc.cod.edu/cgi/viewcontent.cgi?article=1657&context=essai#:~:text=The%20ancient%20people%20believed%20the,%2C%20ghosts%2C%20and%20evil%20spirits.

3. Principles of epidemiology. 20 Dec 2021 [cited 2 Mar 2023]. Available: https://www.cdc.gov/csels/dsepd/ss1978/lesson1/section2.html

4. Fracastoro G. De contagione et contagiosis morbis et curatione. Girolamo Fracastoro, Opera omnia, Venetiis.

5. Pesapane F, Marcelli S, Nazzaro G. Hieronymi Fracastorii: the Italian scientist who described the “French disease.” An Bras Dermatol. 2015;90: 684–686.

6. Rothman KJ, Greenland S, Lash TL, Others. Modern epidemiology. Wolters Kluwer Health/Lippincott Williams & Wilkins Philadelphia; 2008.

7. Snow J. On the mode of communication of cholera . London: John Churchill, 1855. 2020.

8. Bernell S, Howard SW. Use Your Words Carefully: What Is a Chronic Disease? Front Public Health. 2016;4: 159.

9. Ezzati M, Riboli E. Behavioral and dietary risk factors for noncommunicable diseases. N Engl J Med. 2013;369: 954–964.

10. Noncommunicable diseases. [cited 15 Mar 2023]. Available: https://www.who.int/health-topics/noncommunicable-diseases#tab=tab_1/chp/en/

11. Doll R, Hill AB. Smoking and carcinoma of the lung; preliminary report. Br Med J. 1950;2: 739–748.

12. Doll R, Peto R, Wheatley K, Gray R, Sutherland I. Mortality in relation to smoking: 40 years’ observations on male British doctors. BMJ. 1994;309: 901–911.

13. Tenny S, Kerndt CC, Hoffman MR. Case Control Studies. StatPearls. Treasure Island (FL): StatPearls Publishing; 2022.

14. Barrett D, Noble H. What are cohort studies? Evid Based Nurs. 2019;22: 95–96.

15. Savitz DA, Poole C, Miller WC. Reassessing the role of epidemiology in public health. Am J Public Health. 1999;89: 1158–1161.

16. Marathe M, Ramakrishnan N. Recent Advances in Computational Epidemiology. IEEE Intell Syst. 2013;28: 96–101.

17. Gorder PF. Computational Epidemiology. Comput Sci Eng. 2010;12: 4–6.

18. Oster RA. An Examination of Five Statistical Software Packages for Epidemiology. Am Stat. 1998;52: 267–280.

19. Population and health: An introduction to epidemiology. In: PRB [Internet]. [cited 23 Mar 2023]. Available: https://www.prb.org/resources/population-and-health-an-introduction-to-epidemiology/

20. Kivits J, Ricci L, Minary L. Interdisciplinary research in public health: the “why” and the “how.” J Epidemiol Community Health. 2019;73: 1061–1062.

21. Carter-Pokras OD, Offutt-Powell TN, Kaufman JS, Giles WH, Mays VM. Epidemiology, policy, and racial/ethnic minority health disparities. Ann Epidemiol. 2012;22: 446–455.

22. National Research Council (US) Committee on Environmental Epidemiology. Introduction. National Academies Press (US); 1991.

23. Barreto ML, Teixeira MG, Carmo EH. Infectious diseases epidemiology. J Epidemiol Community Health. 2006;60: 192–195.

24. Mattiuzzi C, Lippi G. Current Cancer Epidemiology. J Epidemiol Glob Health. 2019;9: 217–222.

25. Forouhi NG, Wareham NJ. Epidemiology of diabetes. Medicine . 2014;42: 698–702.

26. Ibrahim NK. Epidemiologic surveillance for controlling Covid-19 pandemic: types, challenges and implications. J Infect Public Health. 2020;13: 1630–1638.

27. Fagherazzi G, Ravaud P. Digital diabetes: Perspectives for diabetes prevention, management and research. Diabetes & Metabolism. 2019. pp. 322–329. doi:10.1016/j.diabet.2018.08.012

28. Internet and social media users in the world 2023. In: Statista [Internet]. [cited 23 Mar 2023]. Available: https://www.statista.com/statistics/617136/digital-population-worldwide/

29. Digital around the world. In: DataReportal – Global Digital Insights [Internet]. [cited 11 Oct 2023]. Available: https://datareportal.com/global-digital-overview

30. Digital footprints. [cited 23 Mar 2023]. Available: https://www.familylives.org.uk/advice/your-family/online-safety/digital-footprints

31. Salathé M. Digital epidemiology: what is it, and where is it going? Life Sci Soc Policy. 2018;14: 1.

32. Ginsberg J, Mohebbi MH, Patel RS, Brammer L, Smolinski MS, Brilliant L. Detecting influenza epidemics using search engine query data. Nature. 2009;457: 1012–1014.

33. Tyrer S, Heyman B. Sampling in epidemiological research: issues, hazards and pitfalls. BJPsych Bull. 2016;40: 57–60.

34. Martínez-Mesa J, González-Chica DA, Duquia RP, Bonamigo RR, Bastos JL. Sampling: how to select participants in my research study? An Bras Dermatol. 2016;91: 326–330.

35. Hargittai E. Potential Biases in Big Data: Omitted Voices on Social Media. Soc Sci Comput Rev. 2020;38: 10–24.

36. Bidargaddi N, Musiat P, Makinen V-P, Ermes M, Schrader G, Licinio J. Digital

footprints: facilitating large-scale environmental psychiatric research in naturalistic settings through data from everyday technologies. Mol Psychiatry. 2017;22: 164.

37. May C, Montori VM, Mair FS. We need minimally disruptive medicine. BMJ. 2009;339: b2803.

38. Leppin AL, Montori VM, Gionfriddo MR. Minimally Disruptive Medicine: A Pragmatically Comprehensive Model for Delivering Care to Patients with Multiple Chronic Conditions. Healthcare (Basel). 2015;3: 50–63.

39. Naidoo N, Nguyen VT, Ravaud P, Young B, Amiel P, Schanté D, et al. The research burden of randomized controlled trial participation: a systematic thematic synthesis of qualitative evidence. BMC Med. 2020;18: 6.

40. Tandoc EC, Eng N. Climate change communication on Facebook, Twitter, Sina Weibo, and other social media platforms. Oxford research encyclopedia of climate. doi:10.1093/acrefore/9780190228620.001.0001/acrefore-9780190228620-e-361

41. Gupta DB, Gupta DM. Social media and freedom of speech and expression: Legal challenges and prospects. IMS Manthan. 2015;10. doi:10.18701/imsmanthan.v10i1.5650

42. Fox S, Associate Director. The social life of health information, 2011. [cited 18 Apr 2023]. Available: https://search.issuelab.org/resources/12475/12475.pdf

43. Bour C, Ahne A, Schmitz S, Perchoux C, Dessenne C, Fagherazzi G. The Use of Social Media for Health Research Purposes: Scoping Review. J Med Internet Res. 2021;23: e25736.

44. American Diabetes Association. Diagnosis and classification of diabetes mellitus. Diabetes Care. 2012;35 Suppl 1: S64–71.

45. The top 10 causes of death. [cited 19 Jul 2023]. Available: https://www.who.int/news-room/fact-sheets/detail/the-top-10-causes-of-death

46. Complications of diabetes. In: Diabetes UK [Internet]. 8 Aug 2018 [cited 18 Apr 2023]. Available: https://www.diabetes.org.uk/guide-to-diabetes/complications

47. Roep BO, Thomaidou S, van Tienhoven R, Zaldumbide A. Type 1 diabetes mellitus as a disease of the β-cell (do not blame the immune system?). Nat Rev Endocrinol. 2021;17: 150–161.

48. Mobasseri M, Shirmohammadi M, Amiri T, Vahed N, Hosseini Fard H, Ghojazadeh M. Prevalence and incidence of type 1 diabetes in the world: a systematic review and meta-analysis. Health Promot Perspect. 2020;10: 98–115.

49. DiMeglio LA, Evans-Molina C, Oram RA. Type 1 diabetes. Lancet. 2018;391: 2449–2462.

50. Abela AG, Fava S. Why is the Incidence of Type 1 Diabetes Increasing? Curr Diabetes Rev. 2021;17: e030521193110.

51. Gregory GA, Robinson TIG, Linklater SE, Wang F, Colagiuri S, de Beaufort C, et al. Global incidence, prevalence, and mortality of type 1 diabetes in 2021 with projection to 2040: a modelling study. Lancet Diabetes Endocrinol. 2022;10: 741–760.

52. Diabetic ketoacidosis. In: Mayo Clinic [Internet]. 6 Oct 2022 [cited 21 Apr 2023]. Available: https://www.mayoclinic.org/diseases-conditions/diabetic-ketoacidosis/symptoms-causes/syc-20371551

53. Mean fasting blood glucose. [cited 21 Apr 2023]. Available: https://www.who.int/data/gho/indicator-metadata-registry/imr-details/2380

54. Glucose tolerance tests: What exactly do they involve? Institute for Quality and Efficiency in Health Care (IQWiG); 2020.

55. Designed S, developed by bka interactive ltd, Auckland. HbA1c testing. In: Health Navigator New Zealand [Internet]. [cited 21 Apr 2023]. Available: https://www.healthnavigator.org.nz/health-a-z/h/hba1c-testing/

56. Juneja R, Hirsch IB, Naik RG, Brooks-Worrell BM, Greenbaum CJ, Palmer JP. Islet cell antibodies and glutamic acid decarboxylase antibodies, but not the clinical phenotype, help to identify type 1(1/2) diabetes in patients presenting with type 2 diabetes. Metabolism. 2001;50: 1008–1013.

57. Subramanian S, Baidal D. The Management of Type 1 Diabetes. MDText.com, Inc.; 2021.

58. Shah RB, Patel M, Maahs DM, Shah VN. Insulin delivery methods: Past, present and future. Int J Pharm Investig. 2016;6: 1–9.

59. Gonder-Frederick L. Lifestyle modifications in the management of type 1 diabetes: still relevant after all these years? Diabetes Technol Ther. 2014;16: 695–698.

60. Rickels MR, Peleckis AJ, Dalton-Bakes C, Naji JR, Ran NA, Nguyen H-L, et al. Continuous Glucose Monitoring for Hypoglycemia Avoidance and Glucose Counterregulation in Long-Standing Type 1 Diabetes. J Clin Endocrinol Metab. 2018;103: 105–114.

61. Goyal R, Jialal I. Diabetes Mellitus Type 2. StatPearls Publishing; 2022.

62. CDC. Type 2 Diabetes. In: Centers for Disease Control and Prevention [Internet]. 14 Feb 2023 [cited 21 Apr 2023]. Available: https://www.cdc.gov/diabetes/basics/type2.html

63. Khan MAB, Hashim MJ, King JK, Govender RD, Mustafa H, Al Kaabi J. Epidemiology of Type 2 Diabetes - Global Burden of Disease and Forecasted Trends. J Epidemiol Glob Health. 2020;10: 107–111.

64. Kolb H, Martin S. Environmental/lifestyle factors in the pathogenesis and prevention of type 2 diabetes. BMC Med. 2017;15: 131.

65. Diabetes symptoms: When diabetes symptoms are a concern. In: Mayo Clinic [Internet]. 3 Jun 2021 [cited 21 Apr 2023]. Available: https://www.mayoclinic.org/diseases-conditions/diabetes/in-depth/diabetes-symptoms/art-20044248

66. Gholap NN, Davies MJ, Mostafa SA, Khunti K. Diagnosing type 2 diabetes and identifying high-risk individuals using the new glycated haemoglobin (HbA1c) criteria. Br J Gen Pract. 2013;63: e165–7.

67. Asif M. The prevention and control the type-2 diabetes by changing lifestyle and dietary pattern. J Educ Health Promot. 2014;3: 1.

68. CDC. Gestational diabetes. In: Centers for Disease Control and Prevention [Internet]. 14 Feb 2023 [cited 21 Apr 2023]. Available: https://www.cdc.gov/diabetes/basics/gestational.html

69. Gestational diabetes. 9 Apr 2022 [cited 21 Apr 2023]. Available: https://www.mayoclinic.org/diseases-conditions/gestational-diabetes/diagnosis-treatment/drc-20355345

70. Greene MF, Solomon CG. Gestational diabetes mellitus -- time to treat. The New England journal of medicine. 2005. pp. 2544–2546.

71. Monogenic diabetes (neonatal diabetes mellitus & MODY). In: National Institute of Diabetes and Digestive and Kidney Diseases [Internet]. NIDDK - National Institute of Diabetes and Digestive and Kidney Diseases; 23 Aug 2022 [cited 21 Apr 2023]. Available: https://www.niddk.nih.gov/health-information/diabetes/overview/what-is-diabetes/monogenic-neonatal-mellitus-mody

72. Rajkumar V, Levine SN. Latent Autoimmune Diabetes. StatPearls Publishing; 2022.

73. Blood sugar testing: Why, when and how. In: Mayo Clinic [Internet]. 1 Feb 2022 [cited 21 Apr 2023]. Available: https://www.mayoclinic.org/diseases-conditions/diabetes/in-depth/blood-sugar/art-20046628

74. Greco DS. Diabetes Mellitus in Animals. Nutritional and Therapeutic Interventions for Diabetes and Metabolic Syndrome. Elsevier; 2018. pp. 507–517.

75. Hieronymus L, Geil P. Diabetes basics. Understanding insulin. Diabetes Self Manag. 2005;22: 12, 14, 16, 18, 20–1.

76. Shulman GI. Ectopic fat in insulin resistance, dyslipidemia, and cardiometabolic disease. N Engl J Med. 2014;371: 1131–1141.

77. Hall JE, Hall ME. Guyton and Hall Textbook of Medical Physiology E-Book. Elsevier Health Sciences; 2020.

78. Hirsch LJ, Strauss KW. The Injection Technique Factor: What You Don’t Know or Teach Can Make a Difference. Clin Diabetes. 2019;37: 227–233.

79. Nimri R, Nir J, Phillip M. Insulin Pump Therapy. Am J Ther. 2020;27: e30–e41.

80. Cunningham SM, Tanner DA. A Review: The Prospect of Inhaled Insulin Therapy via Vibrating Mesh Technology to Treat Diabetes. Int J Environ Res Public Health. 2020;17. doi:10.3390/ijerph17165795

81. Fagherazzi G. Technologies will not make diabetes disappear: how to integrate the concept of diabetes distress into care. Diabetes Epidemiology and Management. 2023;11: 100140.

82. Baek RN, Tanenbaum ML, Gonzalez JS. Diabetes burden and diabetes distress: the buffering effect of social support. Ann Behav Med. 2014;48: 145–155.

83. Skinner TC, Joensen L, Parkin T. Twenty-five years of diabetes distress research. Diabet Med. 2020;37: 393–400.

84. Berry E, Lockhart S, Davies M, Lindsay JR, Dempster M. Diabetes distress: understanding the hidden struggles of living with diabetes and exploring intervention strategies. Postgrad Med J. 2015;91: 278–283.

85. Grulovic N, Rojnic Kuzman M, Baretic M. Prevalence and predictors of diabetes-related distress in adults with type 1 diabetes. Sci Rep. 2022;12: 15758.

86. Kiriella DA, Islam S, Oridota O, Sohler N, Dessenne C, de Beaufort C, et al. Unraveling the concepts of distress, burnout, and depression in type 1 diabetes: A scoping review. EClinicalMedicine. 2021;40: 101118.

87. Kalra S, Jena BN, Yeravdekar R. Emotional and Psychological Needs of People with Diabetes. Indian J Endocrinol Metab. 2018;22: 696–704.

88. Ahne A, Orchard F, Tannier X, Perchoux C, Balkau B, Pagoto S, et al. Insulin pricing and other major diabetes-related concerns in the USA: a study of 46 407 tweets between 2017 and 2019. BMJ Open Diabetes Res Care. 2020;8. doi:10.1136/bmjdrc-2020-001190

89. Makice K. Twitter API: Up and Running: Learn How to Build Applications with the Twitter API. “O’Reilly Media, Inc.”; 2009.

90. Roesslein J. Tweepy: Twitter for python. URL: https://github com/tweepy/tweepy. 2020;484.

91. Bradshaw S, Brazil E, Chodorow K. MongoDB: The Definitive Guide: Powerful and Scalable Data Storage. “O’Reilly Media, Inc.”; 2019.

92. Bernard J. Python Data Analysis with pandas. In: Bernard J, editor. Python Recipes Handbook: A Problem-Solution Approach. Berkeley, CA: Apress; 2016. pp. 37–48.

93. Deep-translator. In: PyPI [Internet]. [cited 15 Feb 2023]. Available: https://pypi.org/project/deep-translator/

94. Hutto C, Gilbert E. VADER: A Parsimonious Rule-Based Model for Sentiment Analysis of Social Media Text. ICWSM. 2014;8: 216–225.

95. Mikolov T, Chen K, Corrado G, Dean J. Efficient Estimation of Word Representations in Vector Space. arXiv [cs.CL]. 2013. Available: http://arxiv.org/abs/1301.3781

96. Pennington J, Socher R, Manning C. GloVe: Global Vectors for Word Representation. Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). Doha, Qatar: Association for Computational Linguistics; 2014. pp. 1532–1543.

97. Peters ME, Neumann M, Iyyer M, Gardner M, Clark C, Lee K, et al. Deep contextualized word representations. arXiv [cs.CL]. 2018. Available: http://arxiv.org/abs/1802.05365

98. Radford A, Narasimhan K, Salimans T, Sutskever I, Others. Improving language understanding by generative pre-training. 2018. Available: https://www.cs.ubc.ca/~amuham01/LING530/papers/radford2018improving.pdf

99. Devlin J, Chang M-W, Lee K, Toutanova K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv [cs.CL]. 2018. Available: http://arxiv.org/abs/1810.04805

100. Sokolova M, Lapalme G. A systematic analysis of performance measures for classification tasks. Inf Process Manag. 2009;45: 427–437.

101. Bishop CM. Pattern Recognition and Machine Learning. Springer New York;

102. Wolf T, Debut L, Sanh V, Chaumond J, Delangue C, Moi A, et al. HuggingFace's Transformers: State-of-the-art Natural Language Processing. arXiv [cs.CL]. 2019. Available: http://arxiv.org/abs/1910.03771

103. Nguyen DQ, Vu T, Nguyen AT. BERTweet: A pre-trained language model for English Tweets. arXiv [cs.CL]. 2020. Available: http://arxiv.org/abs/2005.10200

104. Reimers N, Gurevych I. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. arXiv [cs.CL]. 2019. Available: http://arxiv.org/abs/1908.10084

105. Tausczik YR, Pennebaker JW. The Psychological Meaning of Words: LIWC and Computerized Text Analysis Methods. J Lang Soc Psychol. 2010;29: 24–54.

106. Hastie T, Friedman J, Tibshirani R. The Elements of Statistical Learning. Springer New York;

107. Lloyd SP. Least Squares Quantization in PCM. Technical Report RR-5497, Bell Lab, September 1957. 1957.

108. MacQueen JB. Some Methods for Classification and Analysis of Multivariate Observations. 1966.

109. Harvey K. Encyclopedia of Social Media and Politics. SAGE; 2014.

110. The Social Life of Health Information, 2011. In: Pew Research Center: Internet, Science & Tech [Internet]. 12 May 2011 [cited 29 Jun 2020]. Available: https://www.pewresearch.org/internet/2011/05/12/the-social-life-of-health-information-2011/

111. Bardus M, El Rassi R, Chahrour M, Akl EW, Raslan AS, Meho LI, et al. The Use of Social Media to Increase the Impact of Health Research: Systematic Review. J Med Internet Res. 2020;22: e15607.

112. Clement J. Global digital population as of April 2020. In: www-statista-com [Internet]. Available: https://www-statista-com.proxy.bnl.lu/statistics/617136/digital-population-worldwide/

113. Eysenbach G. Infodemiology and Infoveillance: Framework for an Emerging Set of Public Health Informatics Methods to Analyze Search, Communication and Publication Behavior on the Internet. Journal of Medical Internet Research. 2009. p. e11. doi:10.2196/jmir.1157

114. Frandsen M, Thow M, Ferguson SG. The Effectiveness Of Social Media (Facebook) Compared With More Traditional Advertising Methods for Recruiting Eligible Participants To Health Research Studies: A Randomized, Controlled Clinical Trial. JMIR Research Protocols. 2016. p. e161. doi:10.2196/resprot.5747

115. Wilson E, Kenny A, Dickson-Swift V. Using Blogs as a Qualitative Health Research Tool. International Journal of Qualitative Methods. 2015. p. 160940691561804. doi:10.1177/1609406915618049

116. Schwab-Reese LM, Hovdestad W, Tonmyr L, Fluke J. The potential use of social media and other internet-related data and communications for child maltreatment surveillance and epidemiological research: Scoping review and recommendations. Child Abuse Negl. 2018;85: 187–201.

117. Topolovec-Vranic J, Natarajan K. The Use of Social Media in Recruitment for Medical Research Studies: A Scoping Review. J Med Internet Res. 2016;18: e286.

118. Taylor J, Pagliari C. Comprehensive scoping review of health research using social media data. BMJ Open. 2018;8: e022931.

119. Dol J, Tutelman PR, Chambers CT, Barwick M, Drake EK, Parker JA, et al. Health Researchers' Use of Social Media: Scoping Review. J Med Internet Res. 2019;21. doi:10.2196/13687

120. Bour C, Schmitz S, Ahne A, Perchoux C, Dessenne C, Fagherazzi G. Scoping review protocol on the use of social media for health research purposes. BMJ Open. 2021;11: e040671.

121. Arksey H, O'Malley L. Scoping studies: towards a methodological framework. International Journal of Social Research Methodology. 2005. pp. 19–32. doi:10.1080/1364557032000119616

122. Website. [cited 24 Jun 2020]. Available: Micah DJ Peters, Christina Godfrey, Patricia McInerney, Zachary Munn, Andrea C. Tricco, Hanan Khalil. JBI Reviewer's manual. In: wiki.joannabriggs.org [Internet]. 2019. Available: https://wiki.joannabriggs.org/display/MANUAL/Chapter+11%3A+Scoping+reviews

123. Tricco AC, Lillie E, Zarin W, O'Brien KK, Colquhoun H, Levac D, et al. PRISMA Extension for Scoping Reviews (PRISMA-ScR): Checklist and Explanation. Annals of Internal Medicine. 2018. p. 467. doi:10.7326/m18-0850

124. Gibbons J, Malouf R, Spitzberg B, Martinez L, Appleyard B, Thompson C, et al. Twitter-based measures of neighborhood sentiment as predictors of residential population health. PLoS One. 2019;14: e0219550.

125. Bender JL, Cyr AB, Arbuckle L, Ferris LE. Ethics and Privacy Implications of Using the Internet and Social Media to Recruit Participants for Health Research: A Privacy-by-Design Framework for Online Recruitment. J Med Internet Res. 2017;19: e104.

126. Salathé M, Freifeld CC, Mekaru SR, Tomasulo AF, Brownstein JS. Influenza A (H7N9) and the importance of digital epidemiology. N Engl J Med. 2013;369: 401–404.

127. Kim Y, Huang J, Emery S. Garbage in, Garbage Out: Data Collection, Quality Assessment and Reporting Standards for Social Media Data Use in Health Research, Infodemiology and Digital Disease Detection. J Med Internet Res. 2016;18: e41.

128. Arigo D, Pagoto S, Carter-Harris L, Lillie SE, Nebeker C. Using social media for health research: Methodological and ethical considerations for recruitment and intervention delivery. Digit Health. 2018;4: 2055207618771757.

129. Chapter 11: Scoping reviews - JBI Manual for Evidence Synthesis - JBI GLOBAL WIKI. [cited 7 Jul 2020]. Available: https://wiki.joannabriggs.org/display/MANUAL/Chapter+11%3A+Scoping+reviews

130. CADIMA. [cited 7 Jul 2020]. Available: https://www.cadima.info/

131. Kohl C, McIntosh EJ, Unger S, Haddaway NR, Kecke S, Schiemann J, et al. Online tools supporting the conduct and reporting of systematic reviews and systematic maps: a case study on CADIMA and review of existing tools. Environmental Evidence. 2018. doi:10.1186/s13750-018-0115-5

132. The donut and Altmetric Attention Score. In: Altmetric [Internet]. 9 Jul 2015 [cited 30 Jun 2020]. Available: https://www.altmetric.com/about-our-data/the-donut-and-score/

133. Ford KL, Albritton T, Dunn TA, Crawford K, Neuwirth J, Bull S. Youth Study Recruitment Using Paid Advertising on Instagram, Snapchat, and Facebook: Cross-Sectional Survey Study. JMIR Public Health Surveill. 2019;5: e14080.

134. Colditz JB, Chu K-H, Emery SL, Larkin CR, James AE, Welling J, et al. Toward Real-Time Infoveillance of Twitter Health Messages. Am J Public Health. 2018;108: 1009–1014.

135. Sinnenberg L, DiSilvestro CL, Mancheno C, Dailey K, Tufts C, Buttenheim AM, et al. Twitter as a Potential Data Source for Cardiovascular Disease Research. JAMA Cardiol. 2016;1: 1032–1036.

136. Eichstaedt JC, Schwartz HA, Kern ML, Park G, Labarthe DR, Merchant RM, et al. Psychological language on Twitter predicts county-level heart disease mortality. Psychol Sci. 2015;26: 159–169.

137. Sharma M, Yadav K, Yadav N, Ferdinand KC. Zika virus pandemic—analysis of Facebook as a social media health information platform. American Journal of Infection Control. 2017. pp. 301–302. doi:10.1016/j.ajic.2016.08.022

138. Mikal J, Hurst S, Conway M. Ethical issues in using Twitter for population-level depression monitoring: a qualitative study. BMC Medical Ethics. 2016. doi:10.1186/s12910-016-0105-5

139. Rudra K, Sharma A, Ganguly N, Imran M. Classifying and Summarizing Information from Microblogs During Epidemics. Inf Syst Front. 2018;20: 933–948.

140. Scanfeld D, Scanfeld V, Larson EL. Dissemination of health information through social networks: twitter and antibiotics. Am J Infect Control. 2010;38: 182–188.

141. Watson B, Robinson DHZ, Harker L, Jacob Arriola KR. The Inclusion of African-American Study Participants in Web-Based Research Studies: Viewpoint. Journal of Medical Internet Research. 2016. p. e168. doi:10.2196/jmir.5486

142. Davies B, Kotter M. Lessons From Recruitment to an Internet-Based Survey for Degenerative Cervical Myelopathy: Comparison of Free and Fee-Based Methods. JMIR Research Protocols. 2018. p. e18. doi:10.2196/resprot.6567

143. Wozney L, Turner K, Rose-Davis B, McGrath PJ. Facebook ads to the rescue? Recruiting a hard to reach population into an Internet-based behavioral health intervention trial. Internet Interv. 2019;17: 100246.

144. Yuan P, Bare MG, Johnson MO, Saberi P. Using Online Social Media for Recruitment of Human Immunodeficiency Virus-Positive Participants: A Cross-Sectional Survey. Journal of Medical Internet Research. 2014. p. e117. doi:10.2196/jmir.3229

145. The rise of social media. In: Our World in Data [Internet]. [cited 6 Jul 2020]. Available: https://ourworldindata.org/rise-of-social-media

146. Clement J. Most popular social networks worldwide 2020, by reach. In: Statista [Internet]. 23 Jun 2020 [cited 6 Jul 2020]. Available: https://www-statista-com.proxy.bnl.lu/statistics/274773/global-penetration-of-selected-social-media-sites/

147. Anderson M, Jiang J. Teens, Social Media & Technology 2018. In: pewresearch.org [Internet]. 31 May 2018 [cited 6 Jul 2020]. Available: https://www.pewresearch.org/internet/2018/05/31/teens-social-media-technology-2018/

148. Rait MA, Prochaska JJ, Rubinstein ML. Recruitment of adolescents for a smoking study: use of traditional strategies and social media. Transl Behav Med. 2015;5: 254–259.

149. Schwinn T, Hopkins J, Schinke SP, Liu X. Using Facebook ads with traditional paper mailings to recruit adolescent girls for a clinical trial. Addict Behav. 2017;65: 207–213.

150. Salathé M. Digital Pharmacovigilance and Disease Surveillance: Combining Traditional and Big-Data Systems for Better Public Health. J Infect Dis. 2016;214: S399–S403.

151. Leach LS, Butterworth P, Poyser C, Batterham PJ, Farrer LM. Online Recruitment: Feasibility, Cost, and Representativeness in a Study of Postpartum Women. Journal of Medical Internet Research. 2017. p. e61. doi:10.2196/jmir.5745

152. Chu JL, Snider CE. Use of a social networking web site for recruiting Canadian youth for medical research. J Adolesc Health. 2013;52: 792–794.

153. Cowie JM, Gurney ME. The Use of Facebook Advertising to Recruit Healthy Elderly People for a Clinical Trial: Baseline Metrics. JMIR Res Protoc. 2018;7: e20.

154. Ramo DE, Prochaska JJ. Broad Reach and Targeted Recruitment Using Facebook for an Online Survey of Young Adult Substance Use. Journal of Medical Internet Research. 2012. p. e28. doi:10.2196/jmir.1878

155. Arcia A. Facebook Advertisements for Inexpensive Participant Recruitment Among Women in Early Pregnancy. Health Educ Behav. 2014;41: 237–241.

156. Batterham PJ. Recruitment of mental health survey participants using Internet advertising: content, characteristics and cost effectiveness. Int J Methods Psychiatr Res. 2014;23: 184–191.

157. Garland SM, Wark JD, Tabrizi SN, Jayasinghe Y, Moore E, Fletcher A, et al. P1-S4.32 Recruiting via social networking sites for sexual health research (assessing chlamydia and HPV knowledge). Sexually Transmitted Infections. 2011. pp. A173–A174. doi:10.1136/sextrans-2011-050108.176

158. Pedersen ER, Helmuth ED, Marshall GN, Schell TL, PunKay M, Kurz J. Using Facebook to Recruit Young Adult Veterans: Online Mental Health Research. JMIR Research Protocols. 2015. p. e63. doi:10.2196/resprot.3996

159. Youn SJ, Trinh N-H, Shyu I, Chang T, Fava M, Kvedar J, et al. Using online social media, Facebook, in screening for major depressive disorder among college students. International Journal of Clinical and Health Psychology. 2013. pp. 74–80. doi:10.1016/s1697-2600(13)70010-3

160. Santillana M, Nguyen AT, Dredze M, Paul MJ, Nsoesie EO, Brownstein JS. Combining Search, Social Media, and Traditional Data Sources to Improve Influenza Surveillance. PLoS Comput Biol. 2015;11: e1004513.

161. Golder S, Norman G, Loke YK. Systematic review on the prevalence, frequency and comparative value of adverse events data in social media. Br J Clin Pharmacol. 2015;80: 878.

162. Tricco AC, Zarin W, Lillie E, Jeblee S, Warren R, Khan PA, et al. Utility of social media and crowd-intelligence data for pharmacovigilance: a scoping review. BMC Med Inform Decis Mak. 2018;18: 38.

163. Pappa D, Stergioulas LK. Harnessing social media data for pharmacovigilance: a review of current state of the art, challenges and future directions. International Journal of Data Science and Analytics. 2019. pp. 113–135. doi:10.1007/s41060-019-00175-3

164. Corley CD, Cook DJ, Mikler AR, Singh KP. Using Web and Social Media for Influenza Surveillance. Advances in Experimental Medicine and Biology. 2010. pp. 559–564. doi:10.1007/978-1-4419-5913-3_61

165. Brownstein JS, Freifeld CC, Madoff LC. Digital Disease Detection — Harnessing the Web for Public Health Surveillance. New England Journal of Medicine. 2009. pp. 2153–2157. doi:10.1056/nejmp0900702

166. Robertson C, Yee L. Avian Influenza Risk Surveillance in North America with Online Media. PLoS One. 2016;11: e0165688.

167. Paul MJ, Dredze M, Broniatowski D. Twitter improves influenza forecasting. PLoS Curr. 2014;6. doi:10.1371/currents.outbreaks.90b9ed0f59bae4ccaa683a39865d9117

168. Nikfarjam A, Sarker A, O'Connor K, Ginn R, Gonzalez G. Pharmacovigilance from social media: mining adverse drug reaction mentions using sequence labeling with word embedding cluster features. J Am Med Inform Assoc. 2015;22: 671–681.

169. Rong J, Michalska S, Subramani S, Du J, Wang H. Deep learning for pollen allergy surveillance from twitter in Australia. BMC Med Inform Decis Mak. 2019;19: 208.

170. Conway M, Hu M, Chapman WW. Recent Advances in Using Natural Language Processing to Address Public Health Research Questions Using Social Media and ConsumerGenerated Data. Yearb Med Inform. 2019;28: 208–217.

171. Allen C, Tsou M-H, Aslam A, Nagel A, Gawron J-M. Applying GIS and Machine Learning Methods to Twitter Data for Multiscale Surveillance of Influenza. PLoS One. 2016;11: e0157734.

172. Nguyen QC, Li D, Meng H-W, Kath S, Nsoesie E, Li F, et al. Building a National

Neighborhood Dataset From Geotagged Twitter Data for Indicators of Happiness, Diet, and Physical Activity. JMIR Public Health Surveill. 2016;2: e158.

173. Aramburu MJ, Berlanga R, Lanza I. Social Media Multidimensional Analysis for Intelligent Health Surveillance. Int J Environ Res Public Health. 2020;17. doi:10.3390/ijerph17072289

174. Bravo CA, Hoffman-Goetz L. Tweeting About Prostate and Testicular Cancers: What Are Individuals Saying in Their Discussions About the 2013 Movember Canada Campaign? J Cancer Educ. 2016;31: 559–566.

175. Aleksina A, Akulenka S, Lublóy Á. Success factors of crowdfunding campaigns in medical research: perceptions and reality. Drug Discov Today. 2019;24: 1413–1420.

176. Karmaker Santu SK, Bindschadler V, Zhai CX, Gunter CA. NRF: A Naive Re-identification Framework. NRF: A Naive Re-identification Framework. doi:10.1145/3267323.3268948

177. Hunter RF, Gough A, O'Kane N, McKeown G, Fitzpatrick A, Walker T, et al. Ethical Issues in Social Media Research for Public Health. Am J Public Health. 2018;108. doi:10.2105/AJPH.2017.304249

178. Wekerle C, Vakili N, Stewart SH, Black T. The utility of Twitter as a tool for increasing reach of research on sexual violence. Child Abuse Negl. 2018;85: 220–228.

179. Hammer M. Ethical Considerations When Using Social Media for Research. Oncology Nursing Forum. 2017. pp. 410–412. doi:10.1188/17.onf.410-412

180. McKee R. Ethical issues in using social media for health and health care research. Health Policy. 2013. pp. 298–301. doi:10.1016/j.healthpol.2013.02.006

181. Child RJH, Mentes JC, Pavlish C, Phillips LR. Using Facebook and participant information clips to recruit emergency nurses for research. Nurse Res. 2014;21: 16–21.

182. Hu J, Wong KC, Wang Z. Recruiting migrants for health research through social network sites: an online survey among chinese migrants in australia. JMIR Res Protoc. 2015;4: e46.

183. Social media guidelines for researchers - Staff home, The University of York. [cited 8 Jul 2020]. Available: https://www.york.ac.uk/staff/research/governance/research-policies/social-media/

184. Townsend L, Wallace C. Social Media Research: A Guide to Ethics. Available: https://www.gla.ac.uk/media/Media_487729_smxx.pdf

185. Schillinger D, Chittamuru D, Susana Ramírez A. From "Infodemics" to Health Promotion: A Novel Framework for the Role of Social Media in Public Health. American Journal of Public Health. 2020. pp. e1–e4. doi:10.2105/ajph.2020.305746

186. GUIDELINES FOR RESEARCH INVOLVING SOCIAL MEDIA. In: ryerson.ca [Internet]. [cited 17 Jul 2020]. Available: https://www.ryerson.ca/content/dam/research/documents/ethics/guidelines-for-research-involving-social-media.pdf

187. GUIDELINES FOR OBTAINING CONSENT AND ASSENT. In: ryerson.ca [Internet]. [cited 17 Jul 2020]. Available: https://www.ryerson.ca/content/dam/research/documents/ethics/guidelines-for-obtaining-assent-and-consent.pdf

188. Chapter 11: Scoping reviews - JBI Manual for Evidence Synthesis - JBI GLOBAL WIKI. [cited 7 Jul 2020]. Available: https://wiki.joannabriggs.org/display/MANUAL/Chapter+11%3A+Scoping+reviews

189. Hessel F. Burden of Disease. Encyclopedia of Public Health. Springer, Dordrecht; 2008. pp. 94–96.

190. Organization WH, Others. Burden of disease: what is it and why is it important for safer food. Who int. 2019.

191. Measures of disease burden (event-based and time-based) and population attributable risks including identification of comparison groups appropriate to Public Health. 19 Jun 2010 [cited 20 Oct 2021]. Available: https://www.healthknowledge.org.uk/public-health-textbook/research-methods/1a-epidemiology/measures-disease-burden

192. BridgetChapple. What is diabetes distress and burnout? In: Diabetes UK [Internet]. [cited 17 Feb 2022]. Available: https://www.diabetes.org.uk/guide-to-diabetes/emotions/diabetes-burnout

193. Leveraging Twitter data to understand public sentiment for the COVID‐19 outbreak in Singapore. International Journal of Information Management Data Insights. 2021;1: 100021.

194. Q1 2019 Earnings Report. Twitter; Available: https://s22.q4cdn.com/826641620/files/doc_financials/2019/q1/Q1-2019-Slide-Presentation.pdf

195. Yeung AWK, Kletecka-Pulker M, Eibensteiner F, Plunger P, Völkl-Kernstock S, Willschke H, et al. Implications of Twitter in Health-Related Research: A Landscape Analysis of the Scientific Literature. Front Public Health. 2021;9: 654481.

196. Jordan S, Hovet S, Fung I, Liang H, Fu K-W, Tse Z. Using Twitter for Public Health Surveillance from Monitoring and Prediction to Public Response. Data. 2018. p. 6. doi:10.3390/data4010006

197. Weng J, Lee BS. Event Detection in Twitter. Proceedings of the Fifth International Conference on Weblogs and Social Media, Barcelona, Catalonia, Spain, July 17-21, 2011. 2011 [cited 5 Aug 2021]. Available: http://dx.doi.org/

198. Kreider KE. Diabetes Distress or Major Depressive Disorder? A Practical Approach to Diagnosing and Treating Psychological Comorbidities of Diabetes. Diabetes Ther. 2017;8: 1.

199. World Bank Country and Lending Groups – World Bank Data Help Desk. [cited 11 Oct 2021]. Available: https://datahelpdesk.worldbank.org/knowledgebase/articles/906519-world-bank-country-and-lending-groups

200. Steinhaus H, Others. Sur la division des corps matériels en parties. Bull Acad Polon Sci. 1956;1: 801.

201. Jack RE, Garrod OGB, Schyns PG. Dynamic facial expressions of emotion transmit an evolving hierarchy of signals over time. Curr Biol. 2014;24: 187–192.

202. Mulcahy AW, Schwam D, Edenfield N. Comparing Insulin Prices in the United States to Other Countries. Rand Corporation. 2020.

203. Willner S, Whittemore R, Keene D. “Life or death”: Experiences of insulin insecurity among adults with type 1 diabetes in the United States. SSM - Population Health. 2020. p. 100624. doi:10.1016/j.ssmph.2020.100624

204. Lin X, Xu Y, Pan X, Xu J, Ding Y, Sun X, et al. Global, regional, and national burden and trend of diabetes in 195 countries and territories: an analysis from 1990 to 2025. Sci Rep. 2020;10: 1–11.

205. Zimmermann M, Bunn C, Namadingo H, Gray CM, Lwanda J. Experiences of type 2 diabetes in sub-Saharan Africa: a scoping review. Global Health Research and Policy. 2018;3: 1–13.

206. Özcan B, Rutters F, Snoek FJ, Roosendaal M, Sijbrands EJ, Elders PJM, et al. High Diabetes Distress Among Ethnic Minorities Is Not Explained by Metabolic, Cardiovascular, or Lifestyle Factors: Findings From the Dutch Diabetes Pearl Cohort. Diabetes Care. 2018;41: 1854–1861.

207. Gariepy G, Smith KJ, Schmitz N. Diabetes distress and neighborhood characteristics in people with type 2 diabetes. J Psychosom Res. 2013;75: 147–152.

208. Coccaro EF, Lazarus S, Joseph J, Wyne K, Drossos T, Phillipson L, et al. Emotional Regulation and Diabetes Distress in Adults With Type 1 and Type 2 Diabetes. Diabetes Care. 2021;44: 20–25.

209. Richman LS, Kubzansky L, Maselko J, Kawachi I, Choo P, Bauer M. Positive emotion and health: going beyond the negative. Health Psychol. 2005;24: 422–429.

210. Cohort study. [cited 16 Jan 2023]. Available: https://www.iwh.on.ca/what-researchers-mean-by/cohort-study

211. Setia MS. Methodology Series Module 1: Cohort Studies. Indian J Dermatol. 2016;61: 21–25.

212. Song JW, Chung KC. Observational studies: cohort and case-control studies. Plast Reconstr Surg. 2010;126: 2234–2242.

213. Fagherazzi G, Bour C, Ahne A. Emulating a virtual digital cohort study based on social media data as a complementary approach to traditional epidemiology: When, what for, and how? Diabetes Epidemiology and Management. 2022;7: 100085.

214. Aslam S. Twitter by the numbers (2023): Stats, demographics & Fun Facts. In: Omnicore Agency [Internet]. 9 Mar 2023 [cited 2 May 2023]. Available: https://www.omnicoreagency.com/twitter-statistics/

215. Bour C, Ahne A, Aguayo GA, Fischer A, Marcic D, Kayser P, et al. Global Diabetes Burden: Analysis of Regional Differences to Improve Diabetes Care. 2022. doi:10.2139/ssrn.4128868

216. Sinha A, Porter T, Wilson A. The Use of Online Health Forums by Patients With Chronic Cough: Qualitative Study. J Med Internet Res. 2018;20: e19.

217. Finney Rutten LJ, Blake KD, Greenberg-Worisek AJ, Allen SV, Moser RP, Hesse BW. Online Health Information Seeking Among US Adults: Measuring Progress Toward a Healthy People 2020 Objective. Public Health Rep. 2019;134: 617–625.

218. Van, Drake. Python 3 reference manual. Scotts Valley, CA: CreateSpace.

219. datetime — Basic date and time types. In: Python documentation [Internet]. [cited 15 Feb 2023]. Available: https://docs.python.org/3/library/datetime.html

220. Survival regression — lifelines 0.27.3 documentation. [cited 13 Oct 2022]. Available: https://lifelines.readthedocs.io/en/latest/Survival%20Regression.html

221. Spruance SL, Reid JE, Grace M, Samore M. Hazard ratio in clinical trials. Antimicrob Agents Chemother. 2004;48: 2787–2792.

222. Sentence-transformers/all-mpnet-base-v2 · hugging face. [cited 12 Oct 2022]. Available: https://huggingface.co/sentence-transformers/all-mpnet-base-v2

223. Klein AZ, O'Connor K, Levine LD, Gonzalez-Hernandez G. Using Twitter data for cohort studies of drug safety in pregnancy: A proof-of-concept with beta-blockers. bioRxiv. 2022. doi:10.1101/2022.02.23.22271408

224. Tanner AR, Di Cara NH, Maggio V, Thomas R, Boyd A, Sloan L, et al. Epicosm-a framework for linking online social media in epidemiological cohorts. Int J Epidemiol. 2023. doi:10.1093/ije/dyad020

225. Abu Dabrh AM, Gallacher K, Boehmer KR, Hargraves IG, Mair FS. Minimally disruptive medicine: the evidence and conceptual progress supporting a new era of healthcare. J R Coll Physicians Edinb. 2015;45: 114–117.

226. Launders N, Dotsikas K, Marston L, Price G, Osborn DPJ, Hayes JF. The impact of comorbid severe mental illness and common chronic physical health conditions on hospitalisation: A systematic review and meta-analysis. PLoS One. 2022;17: e0272498.

227. Barnes AL, Murphy ME, Fowler CA, Rempfer MV. Health-related quality of life and overall life satisfaction in people with serious mental illness. Schizophr Res Treatment. 2012;2012: 245103.

228. Filipcic I, Simunovic Filipcic I, Ivezic E, Matic K, Tunjic Vukadinovic N, Vuk Pisk S, et al. Chronic physical illnesses in patients with schizophrenia spectrum disorders are independently associated with higher rates of psychiatric rehospitalization; a cross-sectional study in Croatia. Eur Psychiatry. 2017;43: 73–80.

229. Struijs JN, Baan CA, Schellevis FG, Westert GP, van den Bos GAM. Comorbidity in patients with diabetes mellitus: impact on medical health care utilization. BMC Health Serv Res. 2006;6: 84.

230. Bădescu SV, Tătaru C, Kobylinska L, Georgescu EL, Zahiu DM, Zăgrean AM, et al. The association between Diabetes mellitus and Depression. J Med Life. 2016;9: 120–125.

231. Elliott J, Bodinier B, Whitaker M, Delpierre C, Vermeulen R, Tzoulaki I, et al. COVID-19 mortality in the UK Biobank cohort: revisiting and evaluating risk factors. Eur J Epidemiol. 2021;36: 299–309.

232. All of Us Research Program Investigators, Denny JC, Rutter JL, Goldstein DB, Philippakis A, Smoller JW, et al. The “All of Us” Research Program. N Engl J Med. 2019;381: 668–676.

233. Rideout V, Fox S. Digital health practices, social media use, and mental well-being among teens and young adults in the US Hopelab. 2018. /pdf/a-national-survey-by-hopelab-and-well-being-trust ….

234. Whitaker C, Stevelink S, Fear N. The Use of Facebook in Recruiting Participants for Health Research Purposes: A Systematic Review. J Med Internet Res. 2017;19: e290.

235. Signorini A, Segre AM, Polgreen PM. The Use of Twitter to Track Levels of Disease Activity and Public Concern in the U.S. during the Influenza A H1N1 Pandemic. PLoS One. 2011;6: e19467.

236. Yousefinaghani S, Dara R, Poljak Z, Bernardo TM, Sharif S. The Assessment of Twitter’s Potential for Outbreak Detection: Avian Influenza Case Study. Sci Rep. 2019;9: 18147.

237. Ishii H, Shin K, Tosaki T, Haga T, Nakajima Y, Shiraiwa T, et al. Reproducibility and Validity of a Questionnaire Measuring Treatment Burden on Patients with Type 2 Diabetes: Diabetic Treatment Burden Questionnaire (DTBQ). Diabetes Ther. 2018;9: 1001–1019.

238. [No title]. [cited 4 Jul 2023]. Available: https://diabetesjournals.org/care/article/41/6/1299/36487/Insulin-Access-and-Affordability-Working-Group

239. Shrivastava SR, Shrivastava PS, Ramasamy J. Role of self-care in management of diabetes mellitus. J Diabetes Metab Disord. 2013;12: 14.

240. Rariden C. Diabetes Distress: Assessment and Management of the Emotional Aspect of Diabetes Mellitus. J Nurse Pract. 2019;15: 653–656.

241. Elnaggar A, Ta Park V, Lee SJ, Bender M, Siegmund LA, Park LG. Patients’ Use of Social Media for Diabetes Self-Care: Systematic Review. J Med Internet Res. 2020;22: e14209.

242. Euser AM, Zoccali C, Jager KJ, Dekker FW. Cohort studies: prospective versus retrospective. Nephron Clin Pract. 2009;113: c214–7.

243. Vissers PAJ, Falzon L, van de Poll-Franse LV, Pouwer F, Thong MSY. The impact of having both cancer and diabetes on patient-reported outcomes: a systematic review and directions for future research. J Cancer Surviv. 2016;10: 406–415.

244. AlKhathami AD, Alamin MA, Alqahtani AM, Alsaeed WY, AlKhathami MA, Al-Dhafeeri AH. Depression and anxiety among hypertensive and diabetic primary health care patients. Could patients’ perception of their diseases control be used as a screening tool? Saudi Med J. 2017;38: 621–628.

245. Amini F, Khajevand Khoshli A, Asadi J, Bahar A, Najafipour H, Mirzazadeh A. Obesity Mediates the Effect of Past and Current Mental Health on Diabetes Treatment Outcomes. Iran J Public Health. 2022;51: 2608–2618.

246. Frąckiewicz M. The Applications of Chat GPT4 in the Healthcare Industry. In: TS2

SPACE [Internet]. 14 Mar 2023 [cited 22 Aug 2023]. Available: https://ts2.space/en/the-applications-of-chat-gpt4-in-the-healthcare-industry/

247. Exploring the potential of GPT-4 in healthcare: Transforming the future of medical care. In: InnoHEALTH magazine [Internet]. 19 May 2023 [cited 22 Aug 2023]. Available: https://innohealthmagazine.com/2023/research/exploring-the-potential-of-gpt-4-in-healthcare-transforming-the-future-of-medical-care/

248. Deiana G, Dettori M, Arghittu A, Azara A, Gabutti G, Castiglia P. Artificial Intelligence and Public Health: Evaluating ChatGPT Responses to Vaccination Myths and Misconceptions. Vaccines (Basel). 2023;11. doi:10.3390/vaccines11071217

249. Montagna S, Ferretti S, Klopfenstein LC, Florio A, Pengo MF. Data Decentralisation of LLM-Based Chatbot Systems in Chronic Disease Self-Management. Proceedings of the 2023 ACM Conference on Information Technology for Social Good. New York, NY, USA: Association for Computing Machinery; 2023. pp. 205–212.

250. Reece AG, Danforth CM. Instagram photos reveal predictive markers of depression. EPJ Data Science. 2017;6: 1–12.

251. Fagherazzi G, Fischer A, Ismael M, Despotovic V. Voice for Health: The Use of Vocal Biomarkers from Research to Clinical Practice. Digit Biomark. 2021;5: 78–88.

252. Fischer A, Elbeji A, Aguayo G, Fagherazzi G. Recommendations for Successful Implementation of the Use of Vocal Biomarkers for Remote Monitoring of COVID-19 and Long COVID in Clinical Practice and Research. Interact J Med Res. 2022;11: e40655.

253. Ayadi H, Bour C, Fischer A, Ghoniem M, Fagherazzi G. The Long COVID experience from a patient's perspective: a clustering analysis of 27,216 Reddit posts. Front Public Health. 2023;11. doi:10.3389/fpubh.2023.1227807

254. Boettcher N. Studies of Depression and Anxiety Using Reddit as a Data Source: Scoping Review. JMIR Ment Health. 2021;8: e29487.

255. Wu D, Kasson E, Singh AK, Ren Y, Kaiser N, Huang M, et al. Topics and Sentiment Surrounding Vaping on Twitter and Reddit During the 2019 e-Cigarette and Vaping Use-Associated Lung Injury Outbreak: Comparative Study. J Med Internet Res. 2022;24: e39460.

256. Foufi V, Timakum T, Gaudet-Blavignac C, Lovis C, Song M. Mining of Textual Health Information from Reddit: Analysis of Chronic Diseases With Extracted Entities and Their Relations. J Med Internet Res. 2019;21: e12876.

257. Brosch S, de Ferran A-M, Newbould V, Farkas D, Lengsavath M, Tregunno P. Establishing a Framework for the Use of Social Media in Pharmacovigilance in Europe. Drug Saf. 2019;42: 921–930.

Appendix 1. Overview of included studies for the scoping review

Author(s)	Year	Country	Title	Aims	Type of social media	Population	Disease
Norval, C.; Henderson, T.	2019	UK	Automating Dynamic Consent Decisions for the Processing of Social Media Data in Health Research	To identify and discuss a number of real-world implications if Automating Dynamic Consent Decisions were put into practice	Facebook	NR	NR
Brownstein, J. S.; Freifeld, C. C.; Madoff, L. C.	2009	US	Digital disease detection--harnessing the Web for public health surveillance	To describe how the Web can be harnessed for PublicHealth Surveillance	SM in general	NR	NR
Eysenbach, G.	2009	Canada	Infodemiology and infoveillance: framework for an emerging set of public health informatics methods to analyze search, communication and publication behavior on the Internet	To revisit the emerging fields of infodemiology and infoveillance and proposes an expanded framework, introducing some basic metrics such as information prevalence, concept occurrence ratios, and information incidence	SM in general	NR	NR
Scanfeld, D.; Scanfeld, V.; Larson, E. L.	2010	US	Dissemination of health information through social networks: twitter and antibiotics	To review Twitter status updates mentioning “antibiotic(s)” to determine overarching categories and explore evidence of misunderstanding or misuse of antibiotics	Twitter	NR	Drug use disorder
Chew, C.; Eysenbach, G.	2010	Canada	Pandemics in the age of Twitter: content analysis of Tweets during the 2009 H1N1 outbreak	To monitor the use of the terms “H1N1” versus “swine flu” over time, to conduct a content analysis of “tweets” and to validate Twitter as a real-time content, sentiment, and public attention trend-tracking tool.	Twitter	NR	Swine flu

Corley, C. D.; Cook, D. J.; Mikler, A. R.; Singh, K. P.	2010	US	Using Web and social media for influenza surveillance	To evaluate trends in blog posts that discuss influenza	Spinn3r	NR	Influenza
Villagran, M.	2011	US	Methodological diversity to reach patients along the margins, in the shadows, and on the cutting edge	To determine the use of new communication channels to listen to health care consumers and gather data for research purposes	SM in general	NR	NR
Eke, P. I.	2011	US	Using social media for research and public health surveillance	To provide a brief critical assessment of the use of Twitter for public health surveillance and research	SM in general	NR	NR
Garland, S. M.; Wark, J. D.; Tabrizi, S. N.; Jayasinghe, Y.; Moore, E.; Fletcher, A.; Gunasekaran, B.; Ahmed, N.; Fenner, Y.	2011	Australia	Recruiting via social networking sites for sexual health research (assessing chlamydia and HPV knowledge)	To recruit via social networking sites for sexual health research	Facebook	Adolescents; Young adults	Chlamydia; HPV
Murthy, D.; Gross, A.; Oliveira, D.; Soc, Ieee Comp	2011	US	Understanding Cancer-based Networks in Twitter using Social Network Analysis	To discuss the development of methods to visualize networks and information flow on them using real-time data from the social media website Twitter and how these networks influence health outcomes by examining responses to specific health message	Twitter	NR	Cancer
Salathe, M.; Khandelwal, S.	2011	US	Assessing vaccination sentiments with online social media: implications for infectious disease dynamics and control	To measure the spatio-temporal sentiment towards a new vaccine	Twitter	NR	Vaccination

Eysenbach, G.	2011	Canada	Infodemiology and infoveillance tracking online health information and cyber behavior for public health	To develop a proof-of-concept infoveillance system called Infovigil, which can identify, archive, and analyze health-related information from Twitter and other information streams from Internet and social me-dia sources	SM in general	NR	NR
Signorini, A.; Segre, A. M.; Polgreen, P. M.	2011	US	The use of Twitter to track levels of disease activity and public concern in the U.S. during the influenza A H1N1 pandemic	To examine the use of information embedded in the Twitter stream to track rapidly-evolving public sentiment with respect to H1N1 or swine flu, and to track and measure actual disease activity.	Twitter	NR	Influenza A
Fenner, Y.; Garland, S. M.; Moore, E. E.; Jayasinghe, Y.; Fletcher, A.; Tabrizi, S. N.; Gunasekaran, B.; Wark, J. D.	2012	Australia	Web-Based Recruiting for Health Research Using a Social Networking Site: An Exploratory Study	To assess the feasibility of recruiting young females using targeted advertising on the social networking site Facebook	Facebook	Adolescents; Young adults	NR
Salathe, M.; Bengtsson, L.; Bodnar, T. J.; Brewer, D. D.; Brownstein, J. S.; Buckee, C.; Campbell, E. M.; Cattuto, C.; Khandelwal, S.; Mabry, P. L.; Vespignani, A.	2012	US	Digital epidemiology	To describe the potential of digital epidemiology	SM in general	NR	NR
Ramo, D. E.; Prochaska, J. J.	2012	US	Broad reach and targeted recruitment using Facebook for an online survey of young adult substance use	To examine Facebook as a mechanism to reach and survey young adults about tobacco and other substance use	Facebook	Adolescents; Young adults	Substance use

Chunara, R.; Andrews, J. R.; Brownstein, J. S.	2012	US	Social and news media enable estimation of epidemiological patterns early in the 2010 Haitian cholera outbreak	To assess correlation of volume of cholera-related HealthMap news media reports, Twitter postings, and government cholera cases reported in the first 100 days of the 2010 Haitian cholera outbreak	Twitter	NR	Cholera
Balfe, M.; Doyle, F.; Conroy, R.	2012	Ireland	Using Facebook to recruit young adults for qualitative research projects: how difficult is it?	To attempt to contact and recruit individuals with diabetes through Facebook	Facebook	PWSC	Diabetes
Park, B. K.; Calamaro, C.	2013	US	A Systematic Review of Social Networking Sites: Innovative Platforms for Health Research Targeting Adolescents and Young Adults	To review the evidence to determine if social networking sites are effective tools for health research in the adolescent and young adult populations	SM in general	Adolescents ; Young adults	NR
McKee, R.	2013	Germany	Ethical issues in using social media for health and health care research	To summarize the ethical issues to be considered when social media is exploited in healthcare contexts	SM in general	NR	NR
Walker, D. M.	2013	UK	The internet as a medium for health services research. Part 2	To enable readers to make an informed decision about whether online research methods are appropriate for their studies.	Twitter	NR	NR
Centola, D.	2013	US	Social media and the science of health behavior	To demonstrate that the rapid growth of peer-to-peer social media presents an important new resource for addressing these empirical challenges	SM in general	NR	NR
Boicey, C.	2013	US	Innovations in social media: the MappyHealth experience	To review both the opportunities the MappyHealth web application	Twitter	NR	NR

Chu, J. L.; Snider, C. E.	2013	Canada	Use of a Social Networking Web Site for Recruiting Canadian Youth for Medical Research	To describe the effectiveness of facebook advertising in medical research recruitment	Facebook	Adolescents; Young adults	Post-Traumatic stress disorder
Costa, F. F.	2013	US	Social networks, web-based tools and diseases: implications for biomedical research	To review how social networks and other web-based tools are changing the way we approach and track diseases in biomedical research	SM in general	NR	NR
Youn, S. J.; Trinh, N. H.; Shyu, I.; Chang, T.; Fava, M.; Kvedar, J.; Yeung, A.	2013	US	Using online social media, Facebook, in screening for major depressive disorder among college students	To explore the feasibility of using Internet social networking media in an online program for Major Depressive Disorder screening and psychoeducation targeting college students.	Facebook	Students	Depression
Mairs, K.; McNeil, H.; McLeod, J.; Prorok, J. C.; Stolee, P.	2013	Canada	Online strategies to facilitate health-related knowledge transfer: a systematic search and review	To conduct a systematic search and review of the literature on online knowledge translation techniques that foster the interaction between various stakeholders and assisting the sharing of ideas and knowledge within the health field	SM in general	NR	NR
Kapp, J. M.; Peters, C.; Oliver, D. P.	2013	US	Research recruitment using Facebook advertising: big potential, big challenges	To report if Facebook advertising as an exclusive mechanism for recruiting women ages 35–49 years residing in the USA into a health-related research study is effective	Facebook	Adolescents; Young adults	NR
Reaves, A. C.; Bianchi, D. W.	2013	US	The role of social networking sites in medical genetics research	To examine the current role of social networking sites in medical genetics research and potential applications for these sites in	SM in general	NR	NR

				future studies			
Salathe, M.; Freifeld, C. C.; Mekaru, S. R.; Tomasulo, A. F.; Brownstein, J. S.	2013	US	Influenza A (H7N9) and the importance of digital epidemiology	To map the Importance of Digital Epidemiology for Influenza A (H7N9) case	SM in general	NR	Influenza A
Fung, I. C.; Fu, K. W.; Ying, Y.; Schaible, B.; Hao, Y.; Chan, C. H.; Tse, Z. T.	2013	US	Chinese social media reaction to the MERS-CoV and avian influenza A(H7N9) outbreaks	To use Chinese social media data to study the Chinese online community's reaction to the release of official outbreak data from health authorities, namely the outbreaks of MERS-CoV in 2012 and of human infections of avian influenzaA (H7N9) in 2013	Weibo	NR	MERS; Influenza A
Korda H., Itani Z.	2013	US	Harnessing social media for health promotion and behavior change	To summarize current evidence and understanding of using social media for health promotion and to evaluate the effectiveness of various forms of social media and incorporating outcomes research and theory in the design of health promotion programs for social media	SM in general	NR	NR
Gesualdo, F.; Stilo, G.; Agricola, E.; Gonfiantini, M. V.; Pandolfi, E.; Velardi, P.; Tozzi, A. E.	2013	Italy	Influenza-like illness surveillance on Twitter through automated learning of naive language	To develop a minimally trained algorithm that exploits the abundance of health-related web pages to identify all jargon expressions related to a specific technical term	Twitter	NR	Influenza

Broniatowski, D. A.; Paul, M. J.; Dredze, M.	2013	US	National and local influenza surveillance through Twitter: an analysis of the 2012-2013 influenza epidemic	To analyze the performance of this system during the most recent 2012–2013 influenza season and to analyze the performance at multiple levels of geographic granularity, unlike past studies that focused on national or regional surveillance.	Twitter	NR	Influenza A
Kass-Hout, T. A.; Alhinnawi, H.	2013	US	Social media in public health	To determine social media opportunities in public health	SM in general	NR	NR
Park, B. K.; Calamaro, C.	2013	US	A systematic review of social networking sites: innovative platforms for health research targeting adolescents and young adults	To review the evidence to determine if social networking sites are effective tools for health research in the adolescent and young adult populations	SM in general	Adolescents ; Young adults	NR
Kim, E. K.; Seok, J. H.; Oh, J. S.; Lee, H. W.; Kim, K. H.	2013	Korea	Use of hangeul twitter to track and predict human influenza infection	To examine the use of information embedded in the Hangeul Twitter stream to detect rapidly evolving public awareness or concern with respect to influenza transmission	Twitter	NR	Influenza
Ghosh, D. D.; Guha, R.	2013	US	What are we 'tweeting' about obesity? Mapping tweets with Topic Modeling and Geographic Information System	To answer the following questions: How can topic modeling be used to identify relevant public health topics such as obesity on Twitter.com? What are the common obesity related themes? What is the spatial pattern of the themes? What are the research challenges of using large conversational datasets from social networking sites?	Twitter	NR	Obesity

Swirsky, E. S.; Hoop, J. G.; Labott, S.	2014	US	Using social media in research: new ethics for a new meme?	To highlight ethical issues surrounding the use of social media in clinical research	SM in general	NR	NR
Child, R. J.; Mentes, J. C.; Pavlish, C.; Phillips, L. R.	2014	US	Using Facebook and participant information clips to recruit emergency nurses for research	To examine the use of social networking sites in recruiting research participants	Facebook	Emergency nurses	NR
Alshaikh, F.; Ramzan, F.; Rawaf, S.; Majeed, A.	2014	UK	Social network sites as a mode to collect health data: a systematic review	To systematically review the available literature and explore the use of SNS as a mode of collecting data for health research	SM in general	NR	NR
Farnan, J. M.	2014	US	Connectivity and consent: does posting imply participation?	To present two specific issues related to the use of social media and health care research	SM in general	NR	NR
Young, S. D.; Jaganath, D.	2014	US	Feasibility of Using Social Networking Technologies for Health Research Among Men Who Have Sex With Men: A Mixed Methods Study	To assess the feasibility and acceptability of using social networking as a health research platform among men who have sex with men	Facebook; Myspace	Men Who Have Sex With Men	NR
Martin-Sanchez, F.; Verspoor, K.	2014	Australia	Big data in medicine is driving big changes	To summarize current research that takes advantage of “Big Data” in health and biomedical informatics applications	SM in general	NR	NR
Amon, K. L.; Campbell, A. J.; Hawke, C.; Steinbeck, K.	2014	Australia	Facebook as a Recruitment Tool for Adolescent Health Research: A Systematic Review	To conduct a systematic review of the literature on the use of Facebook to recruit adolescents for health research	Facebook	Adolescents; Young adults	NR

Zhang, J.; Zhao, Y. M.; Dimitroff, A.	2014	US	A study on health care consumers' diabetes term usage across identified categories	to investigate health care consumers' diabetes term usage patterns based on Yahoo!Answers social question and answers (QA) forum, identified characteristics and relationships among terms within three pairs of related categories identified from the QA log, and revealed users' diabetes term usage patterns.	Yahoo!Answers social question and answers (QA) forum	NR	Diabetes
Keller, B.; Labrique, A.; Jain, K. M.; Pekosz, A.; Levine, O.	2014	US	Mind the gap: social media engagement by public health researchers	To evaluate the extent to which public health professionals are engaged in digital spaces	SM in general	NR	NR
Shere, M.; Zhao, X. Y.; Koren, G.	2014	Canada	The role of social media in recruiting for clinical trials in pregnancy	To assess the effectiveness of social media as a recruitment tool through the comparison of diverse recruitment techniques in two different phases of the trial	Facebook; Twitter	Pregnant women	Pregnancy
Capurro, D.; Cole, K.; Echavarria, M. I.; Joe, J.; Neogi, T.; Turner, A. M.	2014	Chile	The Use of Social Networking Sites for Public Health Practice and Research: A Systematic Review	To identify the use of social networking sites for public health research and practice and to identify existing knowledge gaps	SM in general	NR	NR
Young, S. D.	2014	US	Behavioral insights on big data: using social media for predicting biomedical outcomes	To provide an overview on how social media data can contribute to the emerging field of 'big data' science, describes current approaches for using social media to monitor and predict health behaviors and disease outbreaks	SM in general	NR	NR

Frandsen, M.; Walters, J.; Ferguson, S. G.	2014	Australia	Exploring the viability of using online social media advertising as a recruitment method for smoking cessation clinical trials	To explore the viability of using social media as a recruitment tool in a clinical research trial.	Facebook	Smokers;Vapers	Smoking
Arcia, A.	2014	US	Facebook Advertisements for Inexpensive Participant Recruitment Among Women in Early Pregnancy	To describe the method by which a national sample of women was successfully recruited	Facebook	Pregnant women	Pregnancy
Young, S. D.; Rivers, C.; Lewis, B.	2014	US	Methods of using real-time social media technologies for detection and remote monitoring of HIV outcomes	To establish methods of using real-time social networking data for HIV prevention by assessing 1) whether geolocated conversations about HIVrisk behaviors can be extracted from social networking data, 2) the prevalence and content of these conversations, and 3) the feasibility of using HIV risk-related real-time social media conversations as a method to detect HIV outcomes	SM in general	NR	HIV
Pawelek, K. A.; Oeldorf-Hirsch, A.; Rong, L.	2014	US	Modeling the impact of twitter on influenza epidemics	To develop a simple mathematical model including the dynamics of “tweets” — short, 140-character Twitter messages that may enhance the awareness of disease, change individual’s behavior, and reduce the transmission of disease among a population during an influenza season	SM in general	NR	Influenza

Batterham, P. J.	2014	Australia	Recruitment of mental health survey participants using Internet advertising: content, characteristics and cost effectiveness	To assess cost-effectiveness , the representativeness and the optimal strategy of online samples of online recruitment	Facebook	Adults	NR
Aslam, A. A.; Tsou, M. H.; Spitzberg, B. H.; An, L.; Gawron, J. M.; Gupta, D. K.; Peddecord, K. M.; Nagel, A. C.; Allen, C.; Yang, J. A.; Lindsay, S.	2014	US	The reliability of tweets as a supplementary method of seasonal influenza surveillance	To improve the correlation of tweets to sentinel-provided influenza-like illness rates by city through filtering and a machine-learning classifier, to observe correlations of tweets for emergency department ILIrates by city, and to explore correlations for tweets to laboratory-confirmed influenza cases in San Diego	Twitter	NR	Influenza
Velasco, E.; Agheneza, T.; Denecke, K.; Kirchner, G.; Eckmanns, T.	2014	Germany	Social media and internet-based data in global systems for public health surveillance: a systematic review	To map public health surveillance into indicator-based surveillance and event-based surveillance and to provide an overview of each	SM in general	NR	NR
Curtis, B. L.	2014	US	Social networking and online recruiting for HIV research: ethical challenges	To describe internet-based HIV/AIDS research recruitment and its ethical challenges, and to outline research participant safeguards and best practices	SM in general	NR	HIV; AIDS
Ruths, D.; Pfeffer, J.	2014	Canada	Social sciences. Social media for large studies of behavior	To highlight issues that are endemic to the study of human behav-ior through large-scale social media datasets and discuss strategies that can be used to address them	SM in general	NR	NR

Martinez, O.; Wu, E.; Shultz, A. Z.; Capote, J.; Lopez Rios, J.; Sandfort, T.; Manusov, J.; Ovejero, H.; Carballo-Dieguez, A.; Chavez Baray, S.; Moya, E.; Lopez Matos, J.; DelaCruz, J. J.; Remien, R. H.; Rhodes, S. D.	2014	US	Still a hard-to-reach population? Using social media to recruit Latino gay couples for an HIV intervention adaptation study	To recruit eligible couples for a study to adapt “Connect ‘n Unite” (an HIV prevention intervention initially created for black gay couples) for Spanish-speaking Latino gay couples living in New York City	Facebook; Instagram; Twitter	Latino	HIV
Paul, M. J.; Dredze, M.; Broniatowski, D.	2014	US	Twitter improves influenza forecasting	To demonstrate that influenza surveillance signals from Twitter significantly improve influenza forecasting	Twitter	NR	Influenza
Velardi, P.; Stilo, G.; Tozzi, A. E.; Gesualdo, F.	2014	Italy	Twitter mining for fine-grained syndromic surveillance	To present a methodology for early detection and analysis of epidemics based on mining Twitter messages	SM in general	NR	Influenza
Yuan, P.; Bare, M. G.; Johnson, M. O.; Saberi, P.	2014	US	Using online social media for recruitment of human immunodeficiency virus-positive participants: a cross-sectional survey	To describe the methodology of a recruitment approach that capitalized on existing online social media venues and other Internet resources in an attempt to overcome some of these barriers to research recruitment and retention.	Facebook; Twitter; Craiglist; Tumblr	NR	HIV
Hartley, D. M.	2014	US	Using social media and internet data for public health surveillance: the importance of talking	To determine how social media and Internet Data can be used for Public Health Surveillance	SM in general	NR	NR

Hays, R.; Daker-White, G.	2015	UK	The care.data consensus? A qualitative analysis of opinions expressed on Twitter	To identify and describe the range of opinions expressed about care data on Twitter for the period during which a delay to this project was announced, and provide insight into the strengths and flaws of the project	Twitter	NR	NR
Meshi, D.; Tamir, D. I.; Heekeren, H. R.	2015	Germany	The Emerging Neuroscience of Social Media	To outline social motives that drive people to use social media, to propose neural systems supporting social media use, and to describe approaches neuroscientists can use to conduct research with social media	SM in general	NR	NR
Pedersen, E. R.; Helmuth, E. D.; Marshall, G. N.; Schell, T. L.; PunKay, M.; Kurz, J.	2015	US	Using Facebook to Recruit Young Adult Veterans: Online Mental Health Research	To recruit a sample of young adult veterans for the first phase of an online alcohol intervention study	Facebook	Adolescents ; Young adults	Drinking
Ramo, D. E.; Thrul, J.; Delucchi, K. L.; Ling, P. M.; Hall, S. M.; Prochaska, J. J.	2015	US	The Tobacco Status Project (TSP): Study protocol for a randomized controlled trial of a Facebook smoking cessation intervention for young adults	To test the efficacy of a stage-based smoking cessation intervention on Facebook for young adults age 18 to 25 on smoking abstinence, reduction in cigarettes smoked, and thoughts about smoking abstinence	Facebook	Adolescents ; Young adults	Smoking
Robinson, B.; Sparks, R.; Power, R.; Cameron, M.	2015	Australia	Social Media Monitoring for Health Indicators	To consider the issue of selection bias inherent in the Twitter data source before population inferences can be made.	Twitter	NR	NR
Cleminson, J.	2015	UK	Child health research and social media	To determine opportunities of social media for child health research	SM in general	NR	NR

Yin, Z. J.; Fabbri, D.; Rosenbloom, S. T.; Malin, B.	2015	US	A Scalable Framework to Detect Personal Health Mentions on Twitter	To develop a scalable framework to detect personal health status mentions on Twitter and assess the extent to which such information is disclosed	Twitter	NR	NR
Toseeb, U.; Inkster, B.	2015	UK	Online social networking sites and mental health research	To provide an account of the theoretical importance of using data generated from social networking sites in mental health research and provide a brief overview of the literature published in this area	SM in general	NR	Depression
Larsen, M. E.; Boonstra, T. W.; Batterham, P. J.; O'Dea, B.; Paris, C.; Christensen, H.	2015	Australia	We Feel: Mapping Emotion on Twitter	To describe the “We Feel” system for analyzing global and regional variations in emotional expression, and report the results of validation against known patterns of variation inmood	Twitter	NR	NR
Rait, M. A.; Prochaska, J. J.; Rubinstein, M. L.	2015	US	Recruitment of adolescents for a smoking study: use of traditional strategies and social media	To examine and compare traditional and Facebook-based recruitment strategies on reach, enrollment, cost, and retention	Facebook	Adolescents; Young adults	Smoking
Kleinsman, J.; Buckley, S.	2015	New-Zealand	Facebook Study: A Little Bit Unethical But Worth It?	To map ethical issues of Facebook studies	SM in general	NR	NR

Hu, J.; Wong, K. C.; Wang, Z. Q.	2015	Australia	Recruiting Migrants for Health Research Through Social Network Sites: An Online Survey Among Chinese Migrants in Australia	To report the process and outcomes of recruiting Chinese migrants through social network sites in Australia and to examine the sample characteristics of online recruitment by comparing the sample which was recruited by an online survey to a sample of Australian Chinese migrants collected by a postal survey	Oursteps; Ozyoyo; Freeoz; Yeeyi; ozchinese; QQ instant message Weibo	Chinese migrants	NR
Wilson, E.; Kenny, A.; Dickson-Swift, V.	2015	Australia	Using Blogs as a Qualitative Health Research Tool: A Scoping Review	To summarize the extent, range, and nature of research activity using blogs	Blogs	NR	NR
Comabella, C. C. I.; Wanat, M.	2015	UK	Using social media in supportive and palliative care research	To provide a comprehensive summary of social media, including its theoretical underpinnings, and recent examples of successful uses of social media in healthcare research	SM in general	NR	NR
Fazzino, T. L.; Rose, G. L.; Pollack, S. M.; Helzer, J. E.	2015	US	Recruiting US and Canadian College Students via Social Media for Participation in a Web-Based Brief Intervention Study	To evaluate the feasibility of recruiting students via free message postings on Facebook and Twitter to participate in a web-based brief intervention study.	Facebook; Twitter	Students	NR
Paris, C.; Christensen, H.; Batterham, P.; O'Dea, B.; Ieee	2015	Australia	Exploring emotions in social media	To show how to explore the emotional state of a population by mining the vast amount of available public social media data in real time	Twitter	NR	NR

Partridge, S. R.; Balestracci, K.; Wong, A. T. Y.; Hebden, L.; McGeechan, K.; Denney-Wilson, E.; Harris, M. F.; Phongsavan, P.; Bauman, A.; Allman-Farinelli, M.	2015	Australia	Effective Strategies to Recruit Young Adults Into the TXT2BFiT mHealth Randomized Controlled Trial for Weight Gain Prevention	To describe the outcomes of strategies used to recruit young adults to a randomized controlled trial, healthy lifestyle mHealth program, TXT2BFiT, for prevention of weight gain	Facebook	Adolescents; Young adults	Nutrition outcomes
Joseph, R. P.; Keller, C.; Adams, M. A.; Ainsworth, B. E.	2015	US	Print versus a culturally-relevant Facebook and text message delivered intervention to promote physical activity in African American women: a randomized pilot trial	To evaluate the potential use of Facebook and text-messaging to deliver culturally-relevant physical activity promotion intervention to African American women	Facebook	Women	Physical activity
Hays, C. A.; Spiers, J. A.; Paterson, B.	2015	Canada	Opportunities and Constraints in Disseminating Qualitative Research in Web 2.0 Virtual Environments	To identify opportunities new digital media presents for knowledge dissemination activities including access to wider audiences with few gatekeeper constraints, new perspectives, and symbiotic relationships between researchers and users	SM in general	NR	NR
Yang, W.; Mu, L.	2015	US	GIS analysis of depression among Twitter users	To apply GIS methods to social media data to provide new per-spectives for public health research	Twitter	NR	Depression

Tempini, N.	2015	UK	Governing PatientsLikeMe: Information Production and Research Through an Open, Distributed, and Data-Based Social Media Network	To explicate how the development of such a data collection architecture requires a continuous exercise of balancing between the conflicting demands of patient engagement, necessary for collecting data in scale, and data semantic context, necessary for effective capture of health phenomena in informative and specific data	PatientsLikeMe	NR	NR
Vayena, E.; Salathe, M.; Madoff, L. C.; Brownstein, J. S.	2015	US	Ethical challenges of big data in public health	To identify some of the key ethical challenges associated with digital disease detection activities and outline a framework for addressing them	SM in general	NR	NR
Lafferty, N. T.; Manca, A.	2015	UK	Perspectives on social media in and as research: A synthetic review	To summarize findings, opinions and discussion about the use of social media in research, including examples from psychiatry and to discuss how the literature can be used to help researchers support the development of personalized research frameworks	SM in general	NR	NR
McGloin, A. F.; Eslami, S.	2015	Ireland	Digital and social media opportunities for dietary behaviour change	To investigate social media opportunities in relation to dietary behaviour change	SM in general	NR	Nutrition outcomes
DeCamp, M.	2015	US	Ethical issues when using social media for health outside professional relationships	To review the major ethical issues arising when social media are used for research, public health, mobile health	SM in general	NR	NR

				applications, and global health			
Adrover, C.; Bodnar, T.; Huang, Z.; Telenti, A.; Salathe, M.	2015	US	Identifying Adverse Effects of HIV Drug Treatment and Associated Sentiments Using Twitter	To assess whether adverse effects of HIV drug treatment and associated sentiments can be determined using publicly available data from social media	Twitter	NR	HIV
Sarker, A.; Ginn, R.; Nikfarjam, A.; O'Connor, K.; Smith, K.; Jayaraman, S.; Upadhaya, T.; Gonzalez, G.	2015	US	Utilizing social media data for pharmacovigilance: A review	To perform a methodical review to characterize the different approaches to adverse drug reaction detection/extraction from social media, and their applicability to pharmacovigilance	SM in general	NR	Adverse drug reaction
Sueki, H.	2015	Japan	The association of suicide-related Twitter use with suicidal behaviour: a cross-sectional study of young internet users in Japan	To examine the association between suicide-related tweets and suicidal behaviour to identify suicidal young people on the Internet	Twitter	Adolescents; Young adults	Suicide
Santillana, M.; Nguyen, A. T.; Dredze, M.; Paul, M. J.; Nsoesie, E. O.; Brownstein, J. S.	2015	US	Combining Search, Social Media, and Traditional Data Sources to Improve Influenza Surveillance	To combine multiple influenza-like illnesses (ILI) activity estimates, generated indepen-dently with each data source, into a single prediction of ILI utilizing machine learning ensem-ble approaches	Twitter	NR	Influenza
Jimeno-Yepes, A.; MacKinlay, A.; Han, B.; Chen, Q.	2015	Australia	Identifying Diseases, Drugs, and Symptoms in Twitter	To study the development of a Twitter data set annotated with relevant medical entities which we have publicly released	Twitter	NR	NR

Zhang, E. X.; Yang, Y.; Di Shang, R.; Simons, J. J.; Quek, B. K.; Yin, X. F.; See, W.; Oh, O. S.; Nandar, K. S.; Ling, V. R.; Chan, P. P.; Wang, Z.; Goh, R. S.; James, L.; Tey, J. S.	2015	Singapore	Leveraging social networking sites for disease surveillance and public sensing: the case of the 2013 avian influenza A(H7N9) outbreak in China	To monitor an avian influenza A (H7N9) outbreak in China and to assess the value of social networking sites in the surveillance of disease outbreaks	Weibo	NR	Influenza A
Nikfarjam, A.; Sarker, A.; O'Connor, K.; Ginn, R.; Gonzalez, G.	2015	US	Pharmacovigilance from social media: mining adverse drug reaction mentions using sequence labeling with word embedding cluster features	To design a machine learning-based approach to extract mentions of adverse drug reactions from highly informal text in social media	DS; Twitter	NR	Adverse drug reaction
Eichstaedt, J. C.; Schwartz, H. A.; Kern, M. L.; Park, G.; Labarthe, D. R.; Merchant, R. M.; Jha, S.; Agrawal, M.; Dziurzynski, L. A.; Sap, M.; Weeg, C.; Larson, E. E.; Ungar, L. H.; Seligman, M. E.	2015	US	Psychological language on Twitter predicts county-level heart disease mortality	To analyze social-media language to identify community-level psychological characteristics associated with mortality from atherosclerotic heart disease	Twitter	NR	Cardiovascular disease
Fung, I. C.; Tse, Z. T.; Fu, K. W.	2015	US	The use of social media in public health surveillance	To overview some of the uses of social media data for public health surveillance and some of the data's strengths and limitations	SM in general	NR	NR
Carceller-Maicas, N.	2016	Spain	Youth, health and social networks Instagram as a research tool for health communication	To map challenges and potential for social health research to reach young people and enrich research ethnographically	Instagram	Adolescents; Young adults	NR

Correia, R. B.; Li, L.; Rocha, L. M.	2016	US	Monitoring potential drug interactions and reactions via network analysis of Instagram user timelines	To determine the potential of Instagram for public health monitoring and surveillance for Drug-Drug Interactions, Adverse Drug reactions, and behavioral pathology at large	Instagram	Instagram users	Drug interactions
Watson, B.; Robinson, D. H. Z.; Harker, L.; Arriola, K. R. J.	2016	US	The Inclusion of African-America n Study Participants in Web-Based Research Studies: Viewpoint	To study how to increase the representation of African-American s in research studies by using the Internet as a recruitment tool and conclude with recommendations that support this goal	SM in general	African Americans	NR
Arayasirikul, S.; Chen, Y. H.; Jin, H.; Wilson, E.	2016	US	A Web 2.0 and Epidemiology Mash-Up: Using Respondent-Dri ven Sampling in Combination with Social Network Site Recruitment to Reach Young Transwomen	To recruit a large, diverse longitudinal cohort of young transwomen using respondent-drive n sampling and to characterize HIV risk and resilience factors in this population	Facebook	Adolescents ; Young adults	HIV
Subasinghe, A. K.; Nguyen, M.; Wark, J. D.; Tabrizi, S. N.; Garland, S. M.	2016	Australia	Targeted Facebook Advertising is a Novel and Effective Method of Recruiting Participants into a Human Papillomavirus Vaccine Effectiveness Study	To determine the feasibility of targeted Facebook advertisements to increase recruitment of unvaccinated women into a human papillomavirus vaccine effectiveness study	Facebook	Adolescents ; Young adults	HPV
Mikal, J.; Hurst, S.; Conway, M.	2016	US	Ethical issues in using Twitter for population-level depression monitoring: a qualitative study	To investigate public opinions on the use of Twitter data for population health monitoring generally, and population mental health monitoring in particular	SM in general	NR	Depression

Khare, R.; Good, B. M.; Leaman, R.; Su, A. I.; Lu, Z.	2016	US	Crowdsourcing in biomedicine: challenges and opportunities	To identify important emerging trends, opportunities and remaining challenges for future crowdsourcing research in biomedicine	SM in general	NR	NR
Kim, Y.; Huang, J. D.; Emery, S.	2016	US	Garbage in, Garbage Out: Data Collection, Quality Assessment and Reporting Standards for Social Media Data Use in Health Research, Infodemiology and Digital Disease Detection	To develop and apply a framework of social media data collection and quality assessment and to propose a reporting standard, which researchers and reviewers may use to evaluate and compare the quality of social data across studies	SM in general	NR	NR
Admon, L.; Haefner, J. K.; Kolenic, G. E.; Chang, T.; Davis, M. M.; Moniz, M. H.	2016	US	Recruiting Pregnant Patients for Survey Research: A Head to Head Comparison of Social Media-Based Versus Clinic-Based Approaches	To compare the feasibility and cost of recruiting pregnant women for survey research using social media-based and clinic-based approaches	Facebook	Pregnant women	Pregnancy
Brodar, K. E.; Hall, M. G.; Butler, E. N.; Parada, H.; Stein-Seroussi, A.; Hanley, S.; Brewer, N. T.	2016	US	Recruiting Diverse Smokers: Enrollment Yields and Cost	To identify effective ways to recruit diverse smokers	Facebook; Craiglist	Smokers; Vapers	Smoking
Schumacher, K. R.; Lee, J. M.	2016	US	Harnessing Social Media for Child Health Research Pediatric Research 2.0	To review both the opportunities and the challenges of using social media to obtain condition-specific, patient-reported data	SM in general	NR	NR

van der Heijden, L.; Piner, S. R.; van de Sande, M. A.	2016	The Netherlands	Pigmented villonodular synovitis: a crowdsourcing study of two hundred and seventy two patients	To ascertain the feasibility of crowdsourcing via Facebook for medical research purposes; by investigating surgical, oncological and functional outcome and quality-of-life in patients with pigmented villonodular synovitis enrolled in a Facebook community	Facebook	NR	Pigmented villonodular synovitis
Winickoff, D. E.; Jamal, L.; Anderson, N. R.	2016	US	New modes of engagement for big data research	To determine new modes of engagement for big data research	SM in general	NR	NR
Gu, L. L.; Skierkowski, D.; Florin, P.; Friend, K.; Ye, Y. J.	2016	US	Facebook, Twitter, & Qr codes: An exploratory trial examining the feasibility of social media mechanisms for sample recruitment	To examine the effectiveness of three social media based recruitment channels for sampling rural adolescent populations for online health research	Facebook; Twitter	Adolescents; Young adults	NR
Devitt, P.	2016	UK	How to use social media to disseminate research findings	To provide guidance to researchers who want to use social media to disseminate research findings	SM in general	NR	NR
Amon, K. L.; Paxton, K.; Klineberg, E.; Riley, L.; Hawke, C.; Steinbeck, K.	2016	Australia	Insights into Facebook Pages: an early adolescent health research study page targeted at parents	To provide a detailed description of the development of the study Facebook Page and present the fan response to the types of posts made on the Page using the Facebook-generated Insights data	Facebook	Parents	NR
Reuter, K.; Ukpolo, F.; Ward, E.; Wilson, M. L.; Angyan, P.	2016	US	Trial Promoter: A Web-Based Tool for Boosting the Promotion of Clinical Research Through Social Media	To develop and test the efficiency of a Web-based tool that automates the generation and distribution of user-friendly social media messages about clinical trials	SM in general	NR	NR

Stephens, S. W.; Williams, C.; Gray, R.; Kerby, J. D.; Wang, H. E.; Bosarge, P. L.	2016	US	Using social media for community consultation and public disclosure in exception from informed consent trials	To describe the experience using social media to facilitate the community consultation and public disclosure process in two trauma resuscitation clinical trials	Facebook	NR	NR
Schootman, M.; Nelson, E. J.; Werner, K.; Shacham, E.; Elliott, M.; Ratnapradipa, K.; Lian, M.; McVay, A.	2016	US	Emerging technologies to measure neighborhood conditions in public health: implications for interventions and next steps	To describe the utility, validity and reli-ability of selected emerging technologies to measure neighborhood conditions for public health applications and to describe next steps for future research and opportunities for interventions	Twitter	NR	NR
Topolovec-Vranic, J.; Natarajan, K.	2016	Canada	The Use of Social Media in Recruitment for Medical Research Studies: A Scoping Review	To determine if social media recruitment is more effective than traditional methods, if social media recruited samples are comparable to those recruited by other methods and if social media is a cost-effective methods	SM in general	NR	NR
Frandsen, M.; Thow, M.; Ferguson, S. G.	2016	Australia	The Effectiveness Of Social Media (Facebook) Compared With More Traditional Advertising Methods for Recruiting Eligible Participants To Health Research Studies: A Randomized, Controlled Clinical Trial	to examine whether visiting the study website prior to being contacted by researchers creates self-screened participants who are more likely to progress through all study phases, to compare conversion percentages and cost effectiveness of each recruitment method at each study phase; and to compare demographic and smoking characteristics of	Facebook	NR	NR

				participants recruited through each strategy to determine if they attract similar samples.			
O'Doherty, K. C.; Christofides, E.; Yen, J.; Bentzen, H. B.; Burke, W.; Hallowell, N.; Koenig, B. A.; Willison, D. J.	2016	Canada	If you build it, they will come: unintended future uses of organized health data collections	To raise awareness of these issues in the hope that they will be discussed more prominently in ELSI (ethical, legal, social implications) and related communities and, ultimately, lead to robust protections.	SM in general	NR	NR
Bravo, C. A.; Hoffman-Goetz, L.	2016	Canada	Tweeting About Prostate and Testicular Cancers: What Are Individuals Saying in Their Discussions About the 2013 Movember Canada Campaign?	To analyze tweets about the 2013 Movember Canada for underlying themes in order understand what those discussions were about	Twitter	NR	Cancer
Conway, M.; O'Connor, D.	2016	US	Social Media, Big Data, and Mental Health: Current Advances and Ethical Implications	To review recent work that utilizes social media “big data” in conjunction with associated technologies like natural language processing and machine learning to address pressing problems in population-level mental health surveillance and research, focusing both on technological advances and core ethical challenges	SM in general	NR	NR
Roccetti, M.; Prandi, C.; Salomoni, P.; Marfia, G.	2016	Italy	Unleashing the true potential of social networks: confirming infliximab medical trials through Facebook posts	To provide implications for both the computer and medical science communities	Facebook	NR	Crohn’s disease; Chronic autoimmune diseases

Spiro, E. S.	2016	US	Research opportunities at the intersection of social media and survey data	To discuss the potential for studies that link social media data with survey data	SM in general	NR	NR
Risson, V.; Saini, D.; Bonzani, I.; Huisman, A.; Olson, M.	2016	Switzerland	Patterns of Treatment Switching in Multiple Sclerosis Therapies in US Patients Active on Social Media: Application of Social Media Content Analysis to Health Outcomes Research	To test the applicability of social media analysis to outcomes research using automated listening combined with filtering and analysis of data by specialists	SM in general	NR	Multiple sclerosis
Salathe, M.	2016	Switzerland	Digital Pharmacovigilance and Disease Surveillance: Combining Traditional and Big-Data Systems for Better Public Health	To map opportunities and challenges of Combining Traditional and Big-Data Systems for Better Public Health	SM in general	NR	NR
Allen, C.; Tsou, M. H.; Aslam, A.; Nagel, A.; Gawron, J. M.	2016	US	Applying GIS and Machine Learning Methods to Twitter Data for Multiscale Surveillance of Influenza	To introduce an improved framework for moni-toring influenza outbreaks using the social media platform Twitter	Twitter	NR	Influenza
Robertson, C.; Yee, L.	2016	Canada	Avian Influenza Risk Surveillance in North America with Online Media	To investigate the use of one social media outlet, Twitter, for surveillance of avian influenza risk in North America	Twitter	NR	Avian influenza
Nguyen, Q. C.; Li, D.; Meng, H. W.; Kath, S.; Nsoesie, E.; Li, F.; Wen, M.	2016	US	Building a National Neighborhood Dataset From Geotagged Twitter Data for Indicators of Happiness, Diet, and Physical Activity	To build, from geotagged Twitter data, a national neighborhood database with area-level indicators of well-being and health behaviors	Twitter	NR	Nutrition outcomes; Physical activity
Allem, J. P.; Ferrara, E.	2016	US	The Importance of Debiasing Social Media Data to Better Understand E-Cigarette-Related Attitudes and Behaviors	To determine the importance of debiasing social media data	SM in general	NR	Vaping

Nguyen, Q. C.; Kath, S.; Meng, H. W.; Li, D.; Smith, K. R.; VanDerslice, J. A.; Wen, M.; Li, F.	2016	US	Leveraging geotagged Twitter data to examine neighborhood happiness, diet, and physical activity	To create neighborhood indicators for happiness, food and physical activity for three large cities: Salt Lake, San Francisco and New York.	Twitter	NR	Nutrition outcomes; Physical activity
Marshall, S. A.; Yang, C. C.; Ping, Q.; Zhao, M.; Avis, N. E.; Ip, E. H.	2016	US	Symptom clusters in women with breast cancer: an analysis of data from social media and a research study	To compare and contrast symptom cluster patterns derived from messages on a breast cancer forum with those from a symptom checklist completed by breast cancer survivors participating in a research study	Medhelp	PWSC	Cancer
Sinnenberg, L.; DiSilvestro, C. L.; Mancheno, C.; Dailey, K.; Tufts, C.; Buttenheim, A. M.; Barg, F.; Ungar, L.; Schwartz, H.; Brown, D.; Asch, D. A.; Merchant, R. M.	2016	US	Twitter as a Potential Data Source for Cardiovascular Disease Research	To characterize the volume and content of tweets related to cardiovascular disease, and the characteristics of Twitter users	Twitter	NR	Cardiovascular disease
Al-Garadi, M. A.; Khan, M. S.; Varathan, K. D.; Mujtaba, G.; Al-Kabsi, A. M.	2016	Malaysia	Using online social networks to track a pandemic: A systematic review	To determine online social networks uses to track a pandemic	SM in general	NR	NR
Braithwaite, S. R.; Giraud-Carrier, C.; West, J.; Barnes, M. D.; Hanson, C. L.	2016	US	Validating Machine Learning Algorithms for Twitter Data Against Established Measures of Suicidality	To validate the use of machine learning algorithms for Twitter data against empirically validated measures of suicidality in the US population	Twitter	NR	Suicide
Helm, C. W.	2017	UK	Social media is essential for research engagement: AGAINST: Likes trumped by dislikes	To map social media opportunities and limitations for health research	SM in general	NR	NR

Mayol, J.; Dziakova, J.	2017	Spain	Value of social media in advancing surgical research	To determine how social media can improve Surgical research	SM in general	NR	NR
Chary, M.; Genes, N.; Giraud-Carrier, C.; Hanson, C.; Nelson, L. S.; Manini, A. F.	2017	US	Epidemiology from Tweets: Estimating Misuse of Prescription Opioids in the USA from Social Media	To estimate misuse of prescription opioids in the USA from Social Media	Twitter	Americans	Drug use disorder
Hammer, M. J.	2017	US	Ethical Considerations When Using Social Media for Research	To summarize the ethical issues to be considered when social media is exploited in healthcare contexts	SM in general	NR	NR
Azer, S. A.	2017	Saudi Arabia	Social Media Channels in Health Care Research and Rising Ethical Issues	To investigate some of the risks inherent in social media research and discuss how researchers should handle challenges related to confidentiality, privacy, and consent when social media tools are used in health-related research	SM in general	NR	NR
Sinnenberg, L.; Buttenheim, A. M.; Padrez, K.; Mancheno, C.; Ungar, L.; Merchant, R. M.	2017	US	Twitter as a Tool for Health Research: A Systematic Review	To systematically review the use of Twitter in health re-search, define a taxonomy to describe Twitter use, and characterize the current state of Twitter in health research	Twitter	NR	NR
Renwick, M. J.; Mossialos, E.	2017	UK	Crowdfunding our health: Economic risks and benefits	To present a typology for crowdfunded health projects and a review of the main economic benefits and risks of crowdfunding in the health market	SM in general	NR	NR

Gelinas, L.; Pierce, R.; Winkler, S.; Cohen, I. G.; Lynch, H. F.; Bierer, B. E.	2017	UK	Using Social Media as a Research Recruitment Tool: Ethical Issues and Recommendations	To defend a non-exceptionalist methodology to assess social media recruitment, to examine respect for privacy and investigator transparency as key norms governing social media recruitment; and to analyze three relatively aspects of social media recruitment	SM in general	NR	NR
Das, R.; Machalek, D. A.; Molesworth, E. G.; Garland, S. M.	2017	Australia	Using Facebook to Recruit Young Australian Men Into a Cross-Sectional Human Papillomavirus Study	To determine the feasibility of using Facebook to recruit young Australian men into human papillomavirus prevalence study	Facebook	Adolescents; Young adults	HPV
Buckarma, E. H.; Thiels, C. A.; Gas, B. L.; Cabrera, D.; Bingener-Casey, J.; Farley, D. R.	2017	US	Influence of Social Media on the Dissemination of a Traditional Surgical Research Article	To determine whether a blog post highlighting the findings of a surgical research article would lead to increased dissemination of the article itself	SM in general	NR	NR
Nelson, K. M.; Ramirez, J. J.; Carey, M. P.	2017	US	Developing Online Recruitment and Retention Methods for HIV Prevention Research Among Adolescent Males Who Are Interested in Sex with Males: Interviews with Adolescent Males	To identify efficient methods to recruit and retain Adolescent males interested in sex with males in online research	SM in general	Adolescents; Young adults	HIV
Motoki, Y.; Miyagi, E.; Taguri, M.; Asai-Sato, M.; Enomoto, T.; Wark, J. D.; Garland, S. M.	2017	Japan	Comparison of Different Recruitment Methods for Sexual and Reproductive Health Research: Social Media-Based Versus Conventional Methods	To determine whether there was a difference in the sexual and reproductive health survey responses of young Japanese women based on recruitment methods: social media–based and conventional methods	Facebook	Adolescents; Young adults	Cancer

Curtis, J. R.; Chen, L.; Higginbotham, P.; Nowell, W. B.; Gal-Levy, R.; Willig, J.; Safford, M.; Coe, J.; O'Hara, K.; Sa'adon, R.	2017	US	Social media for arthritis-related comparative effectiveness and safety research and the impact of direct-to-consumer advertising	to (1) descriptively characterize the demographics of people using social media to discuss rheumatoid arthritis and psoriatic arthritis and psoriasis; (2) to evaluate the suitability of social media as a data source for drug safety research, particularly for the study of recently licensed molecules, and (3) classify the content and timing of the posts that these social media users are contributing, with a particular focus on communication related to newer biologic drugs and small molecules in relation to DTC advertising launch dates	SM in general	NR	Rheumatoid arthritis
Breland, J. Y.; Quintiliani, L. M.; Schneider, K. L.; May, C. N.; Pagoto, S.	2017	US	Social Media as a Tool to Increase the Impact of Public Health Research	To determine opportunities and limitations of social media for health research	SM in general	NR	NR
Bender, J. L.; Cyr, A. B.; Arbuckle, L.; Ferris, L. E.	2017	Canada	Ethics and Privacy Implications of Using the Internet and Social Media to Recruit Participants for Health Research: A Privacy-by-Design Framework for Online Recruitment	To develop a PbD framework for online health research recruitment	Facebook; Twitter; Blogs	NR	NR

Whitaker, C.; Stevelink, S.; Fear, N.	2017	UK	The Use of Facebook in Recruiting Participants for Health Research Purposes: A Systematic Review	To systematically review the literature regarding the current use and success of Facebook To recruit participants for health research purposes	Facebook	NR	NR
Kulanthaivel, A.; Fogel, R.; Jones, J.; Lammert, C.	2017	US	Digital Cohorts Within the Social Mediome: An Approach to Circumvent Conventional Research Challenges?	To determine if social media approach can circumvent conventional Research challenges	SM in general	NR	NR
Mohr, D. C.; Zhang, M.; Schueller, S. M.	2017	US	Personal Sensing: Understanding Mental Health Using Ubiquitous Sensors and Machine Learning	To provide a layered, hierarchical model for translating raw sensor data into markers of behaviors and states related to mental health,	SM in general	NR	NR
Inge, K. J.; Graham, C. W.; McLaughlin, J. W.; Erickson, D.; Wehman, P.; Seward, H. E.	2017	US	Evaluating the effectiveness of Facebook to impact the knowledge of evidence-based employment practices by individuals with traumatic brain injury: A knowledge translation random control study	To compare the effect of a knowledge translation strategy and the use of a secret Facebook group, on the knowledge of evidence-based employment research by individuals with traumatic brain injury	Facebook	NR	Traumatic brain injury
Park, A.; Conway, M.	2017	US	Tracking Health Related Discussions on Reddit for Public Health Applications	To demonstrate social media's potential for public health applications	Reddit	NR	Hemorrhagic Fever; Ebola
Denecke, K.	2017	Switzerland	An ethical assessment model for digital disease detection technologies	To determine the ethical issues to be considered when integrating digital epidemiology with existing practices	SM in general	NR	NR
Leach, L. S.; Butterworth, P.; Poyser, C.; Batterham, P. J.; Farrer, L. M.	2017	Australia	Online Recruitment: Feasibility, Cost, and Representativeness in a Study of Postpartum Women	To investigate the feasibility of recruiting a population of postpartum women online for health research and examine sample	Facebook	Postpartum Women	Postpartum

				representativeness			
Gruebner, O.; Sykora, M.; Lowe, S. R.; Shankardass, K.; Galea, S.; Subramanian, S. V.	2017	Germany	Big data opportunities for social behavioral and mental health research	To determine big data opportunities for social behavioral and mental health research	SM in general	NR	NR
Ade-Ibijola, A.	2017	South Africa	Synthesis of Social Media Profiles Using a Probabilistic Context-Free Grammar	To present a new application of a type of formal grammar — probabilistic/stochastic context-free grammar — in the automatic generation of social media profiles using Facebook as a test case	SM in general	NR	NR
Wilson, R. L.; Usher, K.	2017	Australia	Social media as a recruitment strategy: using Twitter to explore young people's mental health	To discuss the importance of rigorous research designs and to provide an example of a study that demonstrates how mental health researchers, investigating help and support for young people’s mental health, can adapt their traditional recruitment practices and applied this new knowledge to recruitment using social media	Twitter	Adolescents; Young adults	NR
Strekalova, Y. A.; Krieger, J. L.	2017	US	A Picture Really is Worth a Thousand Words: Public Engagement with the National Cancer Institute on Social Media	To determine if facebook is effective for engaging audiences	Facebook	NR	Cancer
Gibson, C. M.	2017	US	The Democratization of Medical Research and Education Through Social Media The Potential and the Peril	To determine opportunities and challenges of social media for Medical Research and Education	SM in general	NR	NR

Schwinn, T.; Hopkins, J.; Schinke, S. P.; Liu, X.	2017	US	Using Facebook ads with traditional paper mailings to recruit adolescent girls for a clinical trial	To detail the use of Facebook ads and traditional paper mailings to enroll 797 adolescent girls for a longitudinal, web-based, drug abuse prevention trial	Facebook	Adolescents; Young adults	Drug use disorder
Ibrahim, A. M.; Lillemoe, K. D.; Klingensmith, M. E.; Dimick, J. B.	2017	US	Visual Abstracts to Disseminate Research on Social Media: A Prospective, Case-control Crossover Study	To compare tweets that included only the title of the article versus tweets that contain the title and a visual abstract	Twitter	NR	NR
Bicquelet, A.	2017	UK	Using online mining techniques to inform formative evaluations: An analysis of YouTube video comments about chronic pain	To argue that exploratory data-mining techniques such as descending hierarchical classification, cluster and correspondence analysis could usefully be employed either as stand-alone or mixed methods in the design of needs assessments on health-related issues	Youtube	NR	Chronic pain
Binder, J. F.; Buglass, S. L.; Betts, L. R.; Underwood, J. D. M.	2017	UK	Online social network data as sociometric markers	To outline in detail the unique information richness of this data type and, in doing so, to support researchers when deciding on ethically appropriate ways of collecting, storing, publishing, and sharing data from online sources	SM in general	NR	NR
Krittanawong, C.; Wang, Z.	2017	US	Mining twitter to understand the smoking cessation barriers	To conduct a data mining analysis of Twitter to assess barriers to smoking cessation	Twitter	NR	Smoking

Roche, B.; Gaillard, B.; Leger, L.; Pelagie-Moutenda, R.; Sochacki, T.; Cazelles, B.; Ledrans, M.; Blateau, A.; Fontenille, D.; Etienne, M.; Simard, F.; Salathe, M.; Yebakima, A.	2017	France	An ecological and digital epidemiology analysis on the role of human behavior on the 2014 Chikungunya outbreak in Martinique	To identify the main drivers of the temporal and spatio-temporal dynamics of the 2014 Chikungunya outbreak in Martinique	Twitter	NR	Chikungunya
Sharma, M.; Yadav, K.; Yadav, N.; Ferdinand, K. C.	2017	US	Zika virus pandemic-analysis of Facebook as a social media health information platform	To examine the effective use of the social mediasite Facebook (Facebook Inc, Menlo Park, CA) as an information source for the Zika virus pandemic	Facebook	NR	Zika
Golder, S.; Ahmed, S.; Norman, G.; Booth, A.	2017	UK	Attitudes Toward the Ethics of Research Using Social Media: A Systematic Review	To ascertain attitudes on the ethical considerations of using social media as a data source for research as expressed by social media users and researchers	SM in general	NR	NR
Birnbaum, M. L.; Ernala, S. K.; Rizvi, A. F.; De Choudhury, M.; Kane, J. M.	2017	US	A Collaborative Approach to Identifying Social Media Markers of Schizophrenia by Employing Machine Learning and Clinical Appraisals	To move from noisy self-reports of schizophrenia on social media to more accurate identification of diagnoses by exploring a human-machine partnered approach, where in computational linguistic analysis of shared content is combined with clinical appraisals	Twitter	NR	Schizophrenia
Sarker, A.; Gonzalez, G.	2017	US	A corpus for mining drug-related knowledge from Twitter chatter: Language models and their utilities	To present to the data science, natural language processing and public health communities an unlabeled corpus and a set of language models	Twitter	NR	Adverse drug reaction

Du, J.; Xu, J.; Song, H. Y.; Tao, C.	2017	US	Leveraging machine learning-based approaches to assess human papillomavirus vaccination sentiment trends with Twitter data	To evaluate a system on a large-scale unannotated tweets corpus and deduce the sentiment labels of those tweets	Twitter	NR	HPV
Nguyen, Q. C.; Meng, H.; Li, D.; Kath, S.; McCullough, M.; Paul, D.; Kanokvimankul, P.; Nguyen, T. X.; Li, F.	2017	US	Social media indicators of the food environment and state health outcomes	To build, from geotagged Twitter and Yelp data, a national food environment database and to test associations between state food environment indicators and health outcomes	Twitter	NR	Nutrition outcomes ; Chronic conditions
Young, S. D.; Yu, W.; Wang, W.	2017	US	Toward Automating HIV Identification: Machine Learning for Rapid Identification of HIV-Related Social Media Data	To test whether four commonly used machine learning methods could learn the patterns associated with HIV risk behavior	Twitter	NR	HIV
Gough, A.; Hunter, R. F.; Ajao, O.; Jurek, A.; McKeown, G.; Hong, J.; Barrett, E.; Ferguson, M.; McElwee, G.; McCarthy, M.; Kee, F.	2017	UK	Tweet for Behavior Change: Using Social Media for the Dissemination of Public Health Messages	To test the feasibility of designing, implementing, and evaluating a social media–enabled intervention for skin cancer prevention	Twitter	NR	Cancer
Choi, I.; Milne, D. N.; Glozier, N.; Peters, D.; Harvey, S. B.; Calvo, R. A.	2017	Australia	Using different Facebook advertisements to recruit men for an online mental health study: Engagement and selection bias	To explore the impact of different Facebook advertisement content for the same study on recruitment rate, engagement, and participant characteristics	Facebook	NR	NR
Sterzing, P. R.; Gartner, R. E.; McGeough, B. L.	2018	USA	Conducting Anonymous, Incentivized, Online Surveys With Sexual and Gender Minority Adolescents: Lessons Learned From a National Polyvictimization Study	to provide guidance to researchers who are designing and implementing anonymous, incentivized, online surveys by examining (a) recruitment and engagement; (b) safety and protection and (c) data integrity	SM in general	Adolescents ; Young adults	Sexuality

Ramo, D. E.; Kaur, M.; Corpuz, E. S.; Satre, D. D.; Delucchi, K.; Brown, S. A.; Prochaska, J. J.	2018	US	Using Facebook to address smoking and heavy drinking in young adults: Protocol for a randomized, controlled trial	To compare the Facebook Tobacco Status Project smoking cessation intervention to an intervention targeting both tobacco use and heavy episodic drinking among young adults who use both substances	Facebook	Adolescents; Young adults	Smoking; Heavy drinking
Young, S. D.; Mercer, N.; Weiss, R. E.; Torrone, E. A.; Aral, S. O.	2018	US	Using social media as a tool to predict syphilis	To assess whether social media could be used to predict syphilis cases in 2013 based on 2012 data	Twitter	Americans	Syphilis
Arigo, D.; Pagoto, S.; Carter-Harris, L.; Lillie, S. E.; Nebeker, C.	2018	US	Using social media for health research: Methodological and ethical considerations for recruitment and intervention delivery	To share recommendations for using social media in recruitment and intervention delivery in health behavior research	SM in general	NR	NR
Timmins, K. A.; Green, M. A.; Radley, D.; Morris, M. A.; Pearce, J.	2018	UK	How has big data contributed to obesity research? A review of the literature	To review the contribution of found data to obesity research and describe the benefits and challenges encountered	Twitter;Facebook;Reddit Foresquare;Instagram;Strava;Online forums	NR	Obesity
Davies, B.; Kotter, M.	2018	UK	Lessons From Recruitment to an Internet-Based Survey for Degenerative Cervical Myelopathy: Comparison of Free and Fee-Based Methods	To compare the efficacy of fee-based advertisement with alternative free recruitment strategies to a Degenerative Cervical Myelopathy Internet health survey	Facebook; Twitter	NR	Degenerative Cervical Myelopathy
Biedermann, N.	2018	Australia	The use of Facebook for virtual asynchronous focus groups in qualitative research	To explore the use of Facebook for recruitment and asynchronous virtual focus group as a data source	Facebook	NR	NR
Mittelstadt, B.; Benzler, J.; Engelmann, L.; Prainsack, B.; Vayena, E.	2018	UK	Is there a duty to participate in digital Mark epidemiology?	To determine whether people have a duty to participate in digital epidemiology	SM in general	NR	NR

Santu, S. K. K.; Bindschadler, V.; Zhai, C. X.; Gunter, C. A.; Acm	2018	US	NRF: A Naive Re-identification Framework	To introduce a general probabilistic reidentification framework that can be instantiated in specific contexts to estimate the probability of compromises based on explicit assumptions.	SM in general	NR	NR
Pang, P. C. I.; Chang, S.; Verspoor, K.; Clavisi, O.	2018	Australia	The Use of Web-Based Technologies in Health Research Participation: Qualitative Study of Consumer and Researcher Experiences	To understand consumers' needs and investigate the opportunities for addressing these needs with Web-based technologies, particularly in the use of Web-based research registers and social networking sites	SM in general	NR	NR
Colditz, J. B.; Chu, K. H.; Emery, S. L.; Larkin, C. R.; James, A. E.; Welling, J.; Primack, B. A.	2018	US	Toward Real-Time Infoveillance of Twitter Health Messages	To provide practical considerations and concrete examples that relate to pilot research on hookah to-bacco smoking, as this trend is both popular and has presented a variety of challenges while working with Twitter data	Twitter	NR	Smoking
Koole, M. A. C.; Kauw, D.; Winter, M. M.; Schuuring, M. J.	2018	The Netherlands	A successful crowdfunding project for eHealth research on grown-up congenital heart disease patients	To fund an eHealthstudy in grown-up congenital heart disease patients and to contemplate on critical success factor	Linkedin; Facebook; Twitter	NR	Cardiovascular disease

Klassen, K. M.; Douglass, C. H.; Brennan, L.; Truby, H.; Lim, M. S. C.	2018	Australia	Social media use for nutrition outcomes in young adults: a mixed-methods systematic review	To describe how young adults use social media in nutrition-related interventions, to evaluate engagement metrics used in social media interventions for nutrition-related outcomes in young adults and to Explore the functions of social media and how these can be leveraged for greatest impact innutrition-related interventions	SM in general	Adolescents; Young adults	Nutrition outcomes
Jha, S. R.; McDonagh, J.; Prichard, R.; Newton, P. J.; Hickman, L. D.; Fung, E.; Macdonald, P. S.; Ferguson, C.	2018	Australia	#Frailty: A snapshot Twitter report on frailty knowledge translation	to explore #Frailty Twitter activity over a six-month period and to provide a snapshot Twitter content analysis of #Frailty usage	Twitter	NR	Frailty
Siegmund, L. A.	2018	US	Social Media: The Next Research Frontier	To provide the clinical nurse specialist with guidance for utilizing social media for research projects and discuss some of the known advantages and obstacles involved in this kind of research	SM in general	NR	NR
Krittanawong, C.; Zhang, H. J.; Aydar, M.; Wang, Z.; Sun, T.	2018	US	Crowdfunding for cardiovascular research	To perform exploratory data analysis of the feasibility of online crowdfunding in cardiovascular research	Facebook; Twitter	NR	NR
Cowie, J. M.; Gurney, M. E.	2018	US	The Use of Facebook Advertising to Recruit Healthy Elderly People for a Clinical Trial: Baseline Metrics	To demonstrate the effectiveness of using targeted advertising on the social networking site Facebook to recruit people aged 60 years and older for volunteer clinical trial participation	Facebook	Elderly People	NR

Ibarra, J. L.; Agas, J. M.; Lee, M.; Pan, J. L.; Buttenheim, A. M.	2018	US	Comparison of Online Survey Recruitment Platforms for Hard-to-Reach Pregnant Smoking Populations: Feasibility Study	To determine the feasibility of recruiting a hard-to-reach population (pregnant smokers) using 4 different Web-based platforms and to compare participants recruited through each platform	Reddit	Pregnant Smokers	Pregnancy; Smoking
Jordan, S. E.; Hovet, S. E.; Fung, I. C. H.; Liang, H.; Fu, K. W.; Tse, Z. T. H.	2018	China	Using Twitter for Public Health Surveillance from Monitoring and Prediction to Public Response	To present the use of mining Twitter data or similar short-text datasets for public health applications	Twitter	NR	Zika; Ebola
Schwab-Reese, L. M.; Hovdestad, W.; Tonmyr, L.; Fluke, J.	2018	US	The potential use of social media and other internet-related data and communications for child maltreatment surveillance and epidemiological research: Scoping review and recommendations	To provide an overview of social media and internet-based methodologies for health research, to report results of evaluation and validation research on these methods, and to highlight studies with potential relevance to child maltreatment research and surveillance	SM in general	NR	Child maltreatment
Gates, A.; Featherstone, R.; Shave, K.; Scott, S. D.; Hartling, L.	2018	Canada	Dissemination of evidence in pediatric emergency medicine: a quantitative descriptive evaluation of a 16-week social media promotion	To develop knowledge products on pediatric emergency medicine topics	Blogs; Twitter	NR	NR
Tricco, A. C.; Zarin, W.; Lillie, E.; Jeblee, S.; Warren, R.; Khan, P. A.; Robson, R.; Pham, B.; Hirst, G.; Straus, S. E.	2018	Canada	Utility of social media and crowd-intelligence data for pharmacovigilance: a scoping review	To characterize the literature the use of conversations in social media as a potential source of data for detecting adverse events related to health products	SM in general	NR	Adverse drug reaction

Hokke, S.; Hackworth, N. J.; Quin, N.; Bennetts, S. K.; Win, H. Y.; Nicholson, J. M.; Zion, L.; Lucke, J.; Keyzer, P.; Crawford, S. B.	2018	Australia	Ethical issues in using the internet to engage participants in family and child research: A scoping review	To identify and integrate evidence on the ethical issues reported when recruiting, retaining and tracing families and children in research online, and to identify ethical guidelines for internet research	SM in general	Families; Children	NR
Cook, C. E.; O'Connell, N. E.; Hall, T.; George, S. Z.; Jull, G.; Wright, A. A.; Girbes, E. L.; Lewis, J.; Hancock, M.	2018	US	Benefits and Threats to Using Social Media for Presenting and Implementing Evidence	To provide a balanced view of benefits and Threats to Using Social Media for Presenting and Implementing Evidence	SM in general	NR	NR
Timberlake, A. T.; Wu, R. T.; Cabrejo, R.; Gabrick, K.; Persing, J. A.	2018	US	Harnessing Social Media to Advance Research in Plastic Surgery	To understand the primary incentives and deterrents for patient par-ticipation in research efforts	Facebook	NR	NR
Wekerle, C.; Vakili, N.; Stewart, S. H.; Black, T.	2018	Canada	The utility of Twitter as a tool for increasing reach of research on sexual violence	To examine the use of social media in child maltreatment research, specifically, the role of social media in knowledge mobilization	Twitter	NR	Child maltreatment
Hunter, R. F.; Gough, A.; O'Kane, N.; McKeown, G.; Fitzpatrick, A.; Walker, T.; McKinley, M.; Lee, M.; Kee, F.	2018	UK	Ethical Issues in Social Media Research for Public Health	To outline the key ethical concerns for public health researchers using SM and discusses how these concerns might best be addressed	SM in general	NR	NR
Sewalk, K. C.; Tuli, G.; Hswen, Y.; Brownstein, J. S.; Hawkins, J. B.	2018	US	Using Twitter to Examine Web-Based Patient Experience Sentiments in the United States: Longitudinal Study	To provide a characterization of patient experience sentiments across the United-States on Twitter over a 4-year period.	Twitter	NR	NR

Zhou, J. Z.; Lemelman, B. T.; Done, N.; Henderson, M. L.; Macmillan, A.; Song, D. H.; Dorafshar, A. H.	2018	US	Social Media and the Dissemination of Research: Insights from the Most Widely Circulated Articles in Plastic Surgery	To quantify the relationship between social media use and the dissemination of research across non traditional channels	SM in general	NR	NR
Gui, X. N.; Kou, Y. B.; Pine, K.; Ladaw, E.; Kim, H.; Suzuki-Gill, E.; Chen, Y. N.; Acm	2018	US	Multidimensional Risk Communication: Public Discourse on Risks during an Emerging Epidemic	To report a qualitative analysis of public perceptions of risks and risk management measures on Reddit during the Zika crisis, an emerging epidemic associated with high uncertainty regarding pathology, epidemiology, and broad consequences	SM in general	NR	Zika
Salathe, M.	2018	Switzerland	Digital epidemiology: what is it, and where is it going?	To provide an outlook of digital epidemiology heading, and offer a broad and a narrow definition of the term	SM in general	NR	NR
Rudra, K.; Sharma, A.; Ganguly, N.; Imran, M.	2018	India	Classifying and Summarizing Information from Microblogs During Epidemics	To build an automatic classification approach useful to categorize tweets into different disease related categories	Twitter	NR	Ebola; MERS
Bernard, R.; Bowsher, G.; Milner, C.; Boyle, P.; Patel, P.; Sullivan, R.	2018	UK	Intelligence and global health: assessing the role of open source and social media intelligence analysis in infectious disease outbreaks	To explore how such tools might be employed in the detection, reporting, and control of outbreaks designated as a ‘threat’ by the global community	SM in general	NR	NR
Xu, X.; Litchman, M. L.; Gee, P. M.; Whatcott, W.; Chacon, L.; Holmes, J.; Srinivasan, S. S.	2018	US	Predicting Prediabetes Through Facebook Postings: Protocol for a Mixed-Methods Study	To investigate the social media behavior of individuals with prediabetes, before and after diagnosis	Facebook	NR	Diabetes

Wakamiya, S.; Kawai, Y.; Aramaki, E.	2018	Japan	Twitter-Based Influenza Detection After Flu Peak via Tweets With Indirect Information: Text Mining Study	To handle indirect information to estimate the trend of the number of influenza patients in each area and each season.	Twitter	NR	Influenza
Akers, L.; Gordon, J. S.	2018	US	Using Facebook for Large-Scale Online Randomized Clinical Trial Recruitment: Effective Advertising Strategies	To share experiences, lessons learned, and recommendations to help researchers design Facebook advertising campaigns	Facebook	NR	NR
Gibbons, J.; Malouf, R.; Spitzberg, B.; Martinez, L.; Appleyard, B.; Thompson, C.; Nara, A.; Tsou, M. H.	2019	US	Twitter-based measures of neighborhood sentiment as predictors of residential population health	To determine how well sentiment predicts self-rated mental health, sleep quality, and heart disease at a census tract level, controlling for neighborhood characteristics and spatial autocorrelation	Twitter	NR	NR
Chowkwanyun, M.	2019	US	Big Data, Large-Scale Text Analysis, and Public Health Research	To familiarize readers with emerging technological trends and methods	SM in general	NR	NR
Reagan, L.; Nowlin, S. Y.; Birdsall, S. B.; Gabbay, J.; Vorderstrasse, A.; Johnson, C.; D'Eramo Melkus, G.	2019	US	Integrative Review of Recruitment of Research Participants Through Facebook	To examine the published evidence concerning the use of Facebook in participant recruitment for adult health research, as compared to other social media, online, and traditional recruitment methods	Facebook	NR	NR
Kimball, S. H.; Hamilton, T.; Benear, E.; Baldwin, J.	2019	US	Determining Emotional Tone and Verbal Behavior in Patients With Tinnitus and Hyperacusis: An Exploratory Mixed-Methods Study	To evaluate the emotional tone and verbal behavior of social media users who self-identified as having tinnitus and;or hyperacusis that caused self-described negative consequences on daily life or health	Facebook	NR	Tinnitus; hyperacusis

McAdam, K.; Warrington, A.; Hughes, A.; Adams, D.; Margham, J.; Vas, C.; Davis, P.; Costigan, S.; Proctor, C.	2019	UK	Use of social media to establish vapers puffing behaviour: Findings and implications for laboratory evaluation of e-cigarette emissions	To generate further information on puffing behaviours of vapers in real-world environment	Youtube	Smokers; Vapers	Vaping
Conway, M.; Hu, M.; Chapman, W. W.	2019	US	Recent Advances in Using Natural Language Processing to Address Public Health Research Questions Using Social Media and ConsumerGenerated Data	To present the utilization of Natural Language Processing for the analysis of social media specifically for public health applications.	SM in general	NR	NR
Wozney, L.; Turner, K.; Rose-Davis, B.; McGrath, P. J.	2019	Canada	Facebook ads to the rescue? Recruiting a hard to reach population into an Internet-based behavioral health intervention trial	To empirically evaluate the impact and cost-effectiveness of paid ads for recruitment into a national trial testing an Internet-based, coached intervention for parents of children with Fetal Alcohol Spectrum Disorders	Facebook	Parents	Fetal Alcohol Spectrum Disorders
Zhan, Y. C.; Etter, J. F.; Leischow, S.; Zeng, D.	2019	US	Electronic cigarette usage patterns: a case study combining survey and social media data	To identify who were social media active e-cigarette users, to compare the use patterns from both survey and social media data for data triangulation, and to jointly use both datasets to conduct a comprehen-sive analysis on e-cigarette future use intentions	Reddit	Smokers; Vapers	Vaping

Shaver, L. G.; Khawer, A.; Yi, Y. Q.; Aubrey-Bassler, K.; Etchegary, H.; Roebothan, B.; Asghari, S.; Wang, P. P.	2019	Canada	Using Facebook Advertising to Recruit Representative Samples: Feasibility Assessment of a Cross-Sectional Survey	To assess Facebook advertising as an economical means of recruiting a representative sample of adults aged 35 to 74 years in Newfoundland and Labrado for a cross-sectional health survey	Facebook	Adults	NR
Graham, S.; Depp, C.; Lee, E. E.; Nebeker, C.; Tu, X.; Kim, H. C.; Jeste, D. V.	2019	USA	Artificial Intelligence for Mental Health and Mental Illnesses: an Overview	To demonstrate AI's potential in mental healthcare	SM in general	NR	Schizophrenia; Depression; Suicide
Bennetts, S. K.; Hokke, S.; Crawford, S.; Hackworth, N. J.; Leach, L. S.; Nguyen, C.; Nicholson, J. M.; Cooklin, A. R.	2019	Australia	Using Paid and Free Facebook Methods to Recruit Australian Parents to an Online Survey: An Evaluation	To determine the feasibility of using Facebook to recruit employed Australian parents to an online survey about managing work and family demands	Facebook	Parents	NR
Jagfeld, G.;	2019	UK	A computational linguistic study of personal recovery in bipolar disorder	To collect and analyze social media data of individuals diagnosed with bipolar disorder with regard to their recovery experience	Twitter; Reddit; Blogs	NR	Bipolar disorder
Jahanbin, K.; Rahmanian, F.; Rahmanian, V.; Jahromi, A. S.	2019	Iran	Application of Twitter and web news mining in infectious disease surveillance systems and prospects for public health	To develop a text mining technique for extracting information about infectious diseases from tweets and news on social media	Twitter	NR	NR
Chan, J. L.; Purohit, H.	2019	US	Challenges to Transforming Unconventional Social Media Data into Actionable Knowledge for Public Health Systems During Disasters	To determine challenges to transforming unconventional data into actionable knowledge for public health systems during disasters	SM in general	NR	NR

Chiauzzi, E.; Wicks, P.	2019	US	Digital Trespass: Ethical and Terms-of-Use Violations by Researchers Accessing Data From an Online Patient Community	To review four cases involving ethical and terms-of-use violations by researchers seeking to conduct social media studies in an online patient research network	SM in general	NR	NR
Grady, K.; Gibson, M.; Bower, P.	2019	UK	Can a 'consent to contact' community help research teams overcome barriers to recruitment? The development and impact of the 'Research for the Future' community	To describe the development of the "Research for the Future" consent to contact community, outline the recruitment of patients to the community, and present data on their participation in research	SM in general	NR	NR
Dol, J.; Tutelman, P. R.; Chambers, C. T.; Barwick, M.; Drake, E. K.; Parker, J. A.; Parker, R.; Benchimol, E. I.; George, R. B.; Witteman, H. O.	2019	Canada	Health Researchers' Use of Social Media: Scoping Review	To explore how social media is used by health researchers professionally, as reported in the literature	SM in general	NR	NR
Hopewell-Kelly, N.; Baillie, J.; Sivell, S.; Harrop, E.; Bowyer, A.; Taylor, S.; Thomas, K.; Newman, A.; Prout, H.; Byrne, A.; Taubert, M.; Nelson, A.	2019	UK	Palliative care research centre's move into social media: constructing a framework for ethical research, a consensus paper	To detail the process in constructing a set of ethical guidelines by which to work	SM in general	NR	NR
McCarthy, E.; Mazza, D.	2019	Australia	Cost and Effectiveness of Using Facebook Advertising to Recruit Young Women for Research: PREFER (Contraceptive Preferences Study) Experience	To evaluate the cost and effectiveness of using Facebook to recruit young women into a Web-based intervention study	Facebook	Adolescents; Young adults	Sexuality

Chancellor, S.; Nitzburg, G.; Hu, A.; Zampieri, F.; De Choudhury, M.; Assoc Comp, Machinery	2019	US	Discovering Alternative Treatments for Opioid Use Recovery Using Social Media	To present the first large-scale social media study of alternative treatments in opioid use disorder recovery, draw-ing on advances in machine learning and computational linguistics	Reddit	NR	Drug use disorder
Wasilewski, M. B.; Stinson, J. N.; Webster, F.; Cameron, J. I.	2019	Canada	Using Twitter to recruit participants for health research: An example from a caregiving study	To describe the nature and extent of study-related tweets, the extent to which they were shared by others, and their potential reach	Twitter	NR	NR
Aparicio-Martinez, P.; Perea-Moreno, A. J.; Martinez-Jimenez, M. P.; Redel-Macias, M. D.; Vaquero-Abellan, M.; Pagliari, C.	2019	Spain	A Bibliometric Analysis of the Health Field Regarding Social Networks and Young People	To map the trends in publications focused on social networks, health, and young people over the last 40 years	SM in general	Adolescents; Young adults	NE
Fagherazzi, G.; Ravaud, P.	2019	France	Digital diabetes: Perspectives for diabetes prevention, management and research	To study how the digitization of diabetes can impact all fields of diabetes – its prevention, management, technology and research – and how it can complement, but not replace, what is usually done in traditional clinical settings.	SM in general	NR	Diabetes
Crawford, S.; Hokke, S.; Nicholson, J. M.; Zion, L.; Lucke, J.; Keyzer, P.; Hackworth, N.	2019	Australia	"It's not black and white" Public health researchers' and ethics committees' perceptions of engaging research participants online	To highlight the key ethical issues of using the internet to recruit, retain and trace participants in public health research, from the perspectives of researchers and human research ethics committee members	SM in general	Researchers	NR

Brieger, K.; Zajac, G. J. M.; Pandit, A.; Foerster, J. R.; Li, K. W.; Annis, A. C.; Schmidt, E. M.; Clark, C. P.; McMorrow, K.; Zhou, W.; Yang, J.; Kwong, A. M.; Boughton, A. P.; Wu, J.; Scheller, C.; Parikh, T.; de la Vega, A.; Brazel, D. M.; Frieser, M.; Rea-Sandin, G.; Fritsche, L. G.; Vrieze, S. I.; Abecasis, G. R.	2019	US	Genes for Good: Engaging the Public in Genetics Research via Social Media	To study the social media to engage a large, diverse participant pool in genetics research and education	Facebook	NR	NR
Lee, H. H.; Hsieh, Y. P.; Murphy, J.; Tidey, J. W.; Savitz, D. A.	2019	US	Health Research Using Facebook to Identify and Recruit Pregnant Women Who Use Electronic Cigarettes: Internet-Based Nonrandomized Pilot Study	To provide information to researchers who seek to recruit participants from rare populations using social media for studies with demanding protocols.	Facebook	Pregnant Smokers	Pregnancy; Smoking
Leach, L. S.; Bennetts, S. K.; Giallo, R.; Cooklin, A. R.	2019	Australia	Recruiting fathers for parenting research using online advertising campaigns: Evidence from an Australian study	To provide information and learnings about recruiting fathers online using social media	Facebook	Fathers	NR
Chen, X.; Lun, Y.; Yan, J.; Hao, T.; Weng, H.	2019	China	Discovering thematic change and evolution of utilizing social media for healthcare research	To discover the thematic change and evolution of utilizing social media for healthcare research field	SM in general	NR	NR

Nelson, E. J.; Loux, T.; Arnold, L. D.; Siddiqui, S. T.; Schootman, M.	2019	US	Obtaining contextually relevant geographic data using Facebook recruitment in public health studies	To report the process of using Facebook recruitment and demonstrate how this strategy can enhance collection of geospatial data to better understand context and spatial patterns of disease	Facebook	NR	NR
Aleksina, A.; Akulenka, S.; Lubloy, A.	2019	Latvia	Success factors of crowdfunding campaigns in medical research: perceptions and reality	To investigate the determinants of successful crowdfunding campaigns in medical research	Twitter	NR	NR
Soragni, A.; Maitra, A.	2019	US	Of scientists and tweets	To determine potential uses of Twitter by scientists	Twitter	NR	NR
Gelinas, L.; Bierer, B. E.	2019	US	Social Media as an Ethical Tool for Retention in Clinical Trials	To describe how social Media can be used as an Ethical Tool for Retention in Clinical Trials	SM in general	NR	NR
Jalal, M.; Wang, K.; Jefferson, S.; Zheng, Y.; Nsoesie, E. O.; Betke, M.; Assoc Comp, Machinery	2019	US	Scraping Social Media Photos Posted in Kenya and Elsewhere to Detect and Analyze Food Types	To create food image datasets from Instagram posts and use it to monitor dietary behavior	Instagram	NR	Nutrition outcomes
Franz, D.; Marsh, H. E.; Chen, J. I.; Teo, A. R.	2019	US	Using Facebook for Qualitative Research: A Brief Primer	To identify opportunities, as well as potential pitfalls, of conducting qualitative research with Facebook users and their activity on Facebook And provide potential options to address each of these issues.	SM in general	NR	NR
Yousefinaghani, S.; Dara, R.; Poljak, Z.; Bernardo, T. M.; Sharif, S.	2019	Canada	The Assessment of Twitter's Potential for Outbreak Detection: Avian Influenza Case Study	To develop a Twitter-based data analysis framework to automatically monitor avian influenza outbreaks in a real-time manner	Twitter	NR	Avian influenza

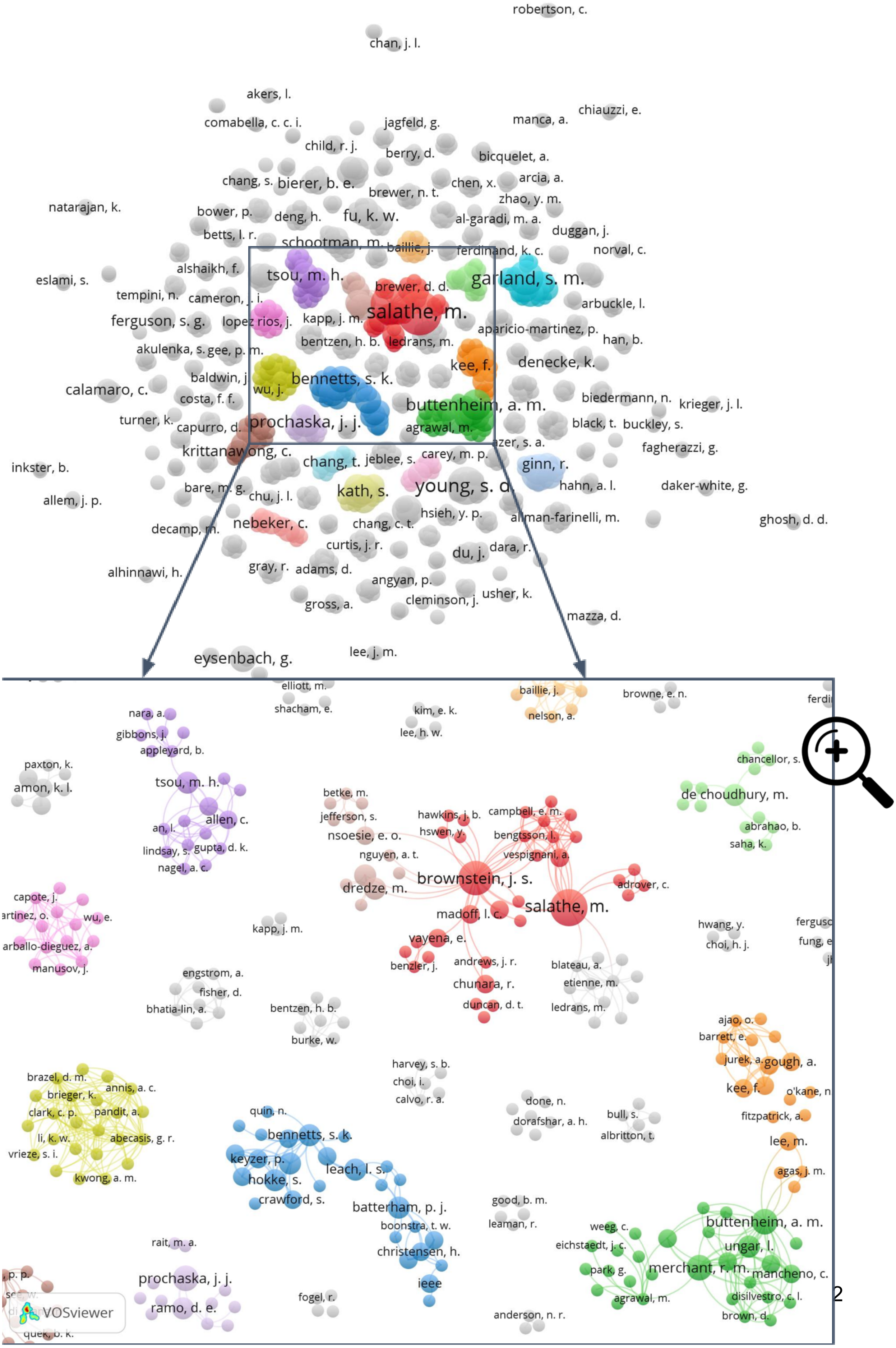
Rong, J.; Michalska, S.; Subramani, S.; Du, J.; Wang, H.	2019	Australia	Deep learning for pollen allergy surveillance from twitter in Australia	To introduce a deep learning-based approach for real-time detection and insights generation about one of the most prevalent chronic conditions in Australia - Pollen allergy.	Twitter	NR	Pollen allergy
Weissenbacher, D.; Sarker, A.; Klein, A.; O'Connor, K.; Magge, A.; Gonzalez-Hernandez, G.	2019	US	Deep neural networks ensemble for detecting medication mentions in tweets	To propose a more advanced method to recognize misspellings or ambiguity with common words	Twitter	NR	NR
Bhatia-Lin, A.; Boon-Dooley, A.; Roberts, M. K.; Pronai, C.; Fisher, D.; Parker, L.; Engstrom, A.; Ingraham, L.; Darnell, D.	2019	US	Ethical and Regulatory Considerations for Using Social Media Platforms to Locate and Track Research Participants	To offer a rubric that can be used in future studies to determine ethical and regulation-consistent use of social media platforms and illustrate the rubric using our study team's experience with Facebook	SM in general	NR	NR
Kamp, K.; Herbell, K.; Magginis, W. H.; Berry, D.; Given, B.	2019	US	Facebook Recruitment and the Protection of Human Subjects	to examine Facebook recruitment in light of the ethical principles of the Belmont Report (respect for persons, beneficence, and justice), to describe ethical challenges that may be faced in Facebook recruitment, and to provide recommendations for researchers interested in adopting this recruitment method	Facebook	NR	NR

Benedict, C.; Hahn, A. L.; Diefenbach, M. A.; Ford, J. S.	2019	US	Recruitment via social media: advantages and potential biases	To compare recruitment strategies (hospital-based v. social media) in enrollment metrics and among enrolled participants, evaluate group differences in patient characteristics and patient reported outcomes	SM in general	Adolescents ; Young adults	NR
Young, S. D.; Padwa, H.; Bonar, E. E.	2019	US	Social Big Data as a Tool for Understanding and Predicting the Impact of Cannabis Legalization	To help developing the knowledge base about critical health and policy questions related to cannabis legalization, but the data sources have limitations	SM in general	NR	Cannabis legalization
Russomanno , J.; Patterson, J. G.; Jabson Tree, J. M.	2019	US	Social Media Recruitment of Marginalized, Hard-to-Reach Populations: Development of Recruitment and Monitoring Guidelines	To investigate the utility, successes, challenges, and positive and negative consequences of using targeted Facebook advertisements as a strategy to recruit transgender and gender nonconforming people into a research study	Facebook	Adolescents ; Young adults	NR
Saha, K.; Sugar, B.; Torous, J.; Abrahao, B.; Kiciman, E.; De Choudhury, M.	2019	US	A Social Media Study on the Effects of Psychiatric Medication Use	To examine psychopathological effects subject to self-reported usage of psychiatric medication.	Twitter	NR	Substance use
Ford, K. L.; Albritton, T.; Dunn, T. A.; Crawford, K.; Neuwirth, J.; Bull, S.	2019	US	Youth Study Recruitment Using Paid Advertising on Instagram, Snapchat, and Facebook: Cross-Sectional Survey Study	To compare recruitment metrics across Instagram, Snapchat,and Facebook for two surveys documenting youth knowledge, attitudes, and behaviors related to retail marijuana in Colorado post legalization	Facebook; Instagram; Snapchat	Adolescents ; Young adults	NR

Hwang, Y.; Kim, H. J.; Choi, H. J.; Lee, J.	2020	Republic of Korea	Exploring Abnormal Behavior Patterns of Online Users With Emotional Eating Behavior: Topic Modeling Study	To analyze the behavior patterns of emotional eaters as the first step to designing a personalized intervention system	Reddit	NR	Nutrition outcomes
Van Swol, L. M.; Chang, C. T.; Kerr, B.; Moreno, M.	2020	US	Linguistic Predictors of Problematic Drinking in Alcohol-related Facebook Posts	to interview participants via Facebook to assess problematic drinking (binge drinking episodes and number of drink)	SM in general	NR	Drinking
Aramburu, M. J.; Berlanga, R.; Lanza, I.	2020	Spain	Social Media Multidimensional Analysis for Intelligent Health Surveillance	To propose a dynamic multidimensional approach to deal with social datastreams	Twitter	NR	NR
Wang, J.; Deng, H.; Liu, B. T.; Hu, A. B.; Liang, J.; Fan, L. Y.; Zheng, X.; Wang, T.; Lei, J. B.	2020	China	Systematic Evaluation of Research Progress on Natural Language Processing in Medicine Over the Past 20 Years: Bibliometric Study on PubMed	To perform a systematic review on the use of natural language processing (NLP) in medical research with the aim of understanding the global progress on NLP research outcomes, content, methods, and study groups involved	SM in general	NR	NR
Rivas, R.; Sadah, S. A.; Guo, Y.; Hristidis, V.	2020	US	Classification of Health-Related Social Media Posts: Evaluation of Post Content-Classifier Models and Analysis of User Demographics	To classify the content (eg, posts that share experiences and seek support) of users who write health-related social media posts and study the effect of user demographics on post content.	Twitter	NR	NR
O'Connor, K.; Sarker, A.; Perrone, J.; Hernandez, G. G.	2020	US	Promoting Reproducible Research for Characterizing Nonmedical Use of Medications Through Data Annotation: Description of a Twitter Corpus and Guidelines	To discuss the creation of an annotated corpus suitable for training supervised classification algorithms for the automatic classification of medication abuse–related chatter.	Twitter	NR	Drug use disorder

Dodson, C. K.; Jackson, D.; Muzny, C. A.; Eaton, E. F.	2020	US	Quantitative evaluation on the challenges and opportunities in the recruitment of young Black men who have sex with men for sexual health research in the southern US	To evaluate the recruitment of young black men who have sex with men, both HIV infected and uninfected, for a sexual health study in Birmingham from 2017 to 2019 and explore alternative patient-centred recruitment methods	SM in general	Adolescents; Young adults	HIV
Mhasawade, V.; Elghafari, A.; Duncan, D. T.; Chunara, R.	2020	US	Role of the Built and Online Social Environments on Expression of Dining on Instagram	To study the role of the built and online social environments in the expression of dining on Instagram in Abu Dhabi	Instagram	NR	Nutrition outcomes
Thomas, V. L.; Chavez, M.; Browne, E. N.; Minnis, A. M.	2020	USA	Instagram as a tool for study engagement and community building among adolescents: A social media pilot study	To evaluate the effectiveness of and engagement with a human-centered, Instagram-based outreach campaign, with a focus on study retention, enhancement of participants' experiences, and increasing community awareness of the study	Instagram	Adolescents; Young adults	NR
Barros, J. M.; Duggan, J.; Rebholz-Schuhmann, D.	2020	Ireland	The Application of Internet-Based Sources for Public Health Surveillance (Infoveillance): Systematic Review	To assess research findings regarding the application of IBSs for public health surveillance (infodemiology or infoveillance)	SM in general	NR	NR
Griffis, H.; Asch, D. A.; Schwartz, H. A.; Ungar, L.; Buttenheim, A. M.; Barg, F. K.; Mitra, N.; Merchant, R. M.	2020	US	Using Social Media to Track Geographic Variability in Language About Diabetes: Analysis of Diabetes-Related Tweets Across the United States	To characterize the language of Twitter users' posts regarding diabetes and describe the correlation of themes with the county-level prevalence of diabetes	Twitter	NR	Diabetes

Appendix 2. Co-authorship networks



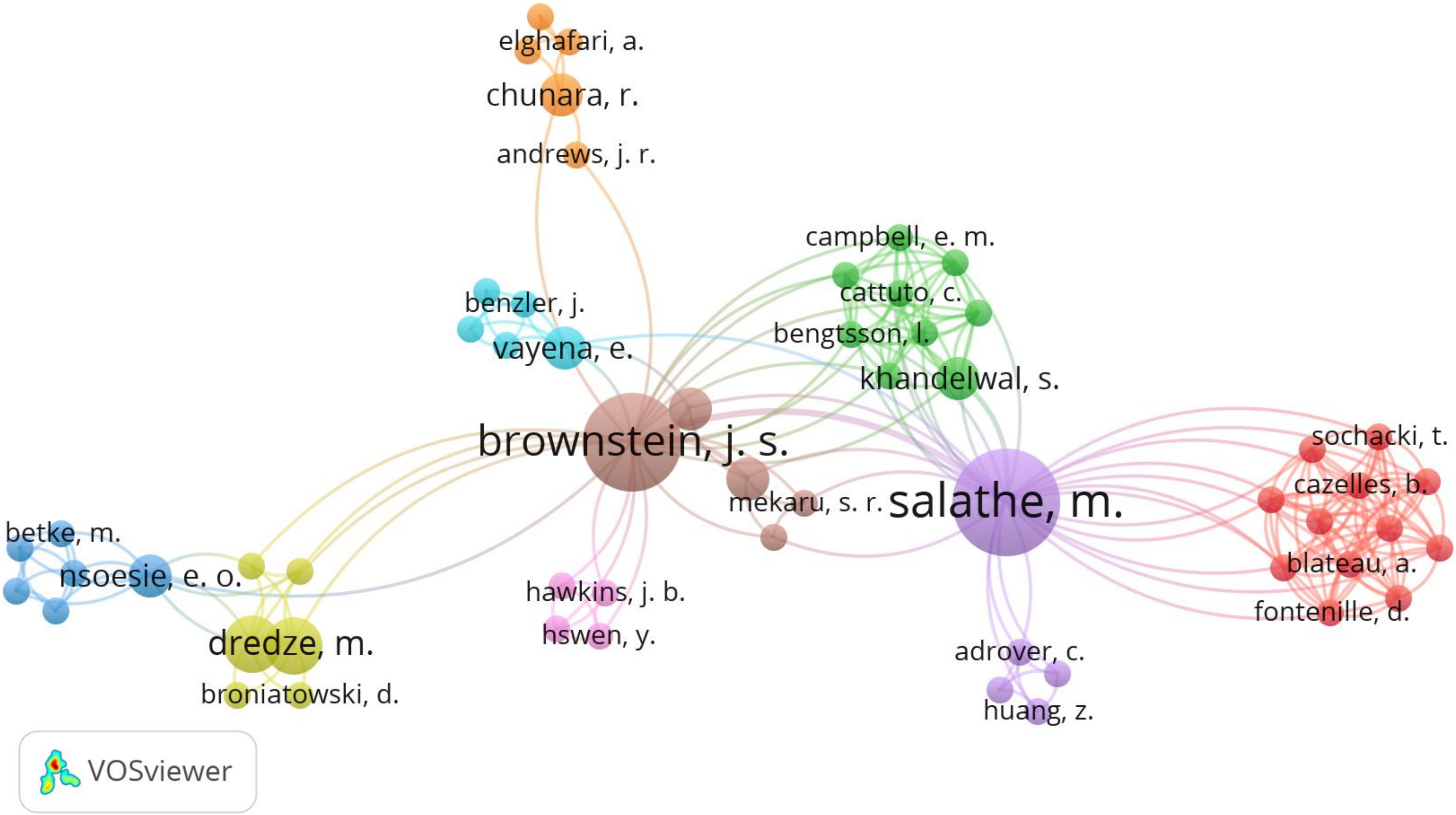


Figure 2. Largest set of connected authors

Appendix 3: Keywords for the collection of diabetes-related tweets

Language	Keywords	Countries covered
Afrikaans	insulien, #insulien, suikersiekte, #suikersiekte, diabeet, #diabeet, Bloedglukose, #blodglucose	South Africa, Namibia
Amharic	ኢንሱሊን, #ኢንሱሊን, የስኳር ህመም, #የስኳር, የስኳር ህመም, #የሱካር በሽታ, የደም ግሉኮስ, #የግሉኮስ	Ethiopia
Arabic	لأنسولين, #الأنسولين, داء السكري, #داءالسكري, مريض بالسكر, #مريض بالسكر, جلوكوز الدم, يكافح السكري, مرض السكر النوع 1, مرض السكر النوع 2	Algeria, Bahrain, Chad, The Comoros, Djibouti, Egypt, Eritrea, Iraq, Jordan, Kuwait, Lebanon, Libya, Mauritania, Morocco, Oman, Palestine, Qatar, Saudi Arabia, Somalia, Sudan, Syria, Tanzania, Tunisia, United Arab Emirates,Yemen
Chinese	胰岛素, #胰岛素, 糖尿病, #胰岛素, 糖尿病, #糖尿病患者, 血糖, #血糖, #2型糖尿病, #1型糖尿病	China, Singapore, Taiwan (Republic of China)
Danish	diabetiske,#diabetiske	Denmark
Dutch	#suikerziekte, bloed glucose, #bloedglucose, #diabetestype1, #diabetestype2	Aruba, Belgium, Curacao, The Netherlands, Sint Maarten, Suriname
English	insulin, #insulin, diabetes, #diabetes, diabetic, #diabetic, #diabeticproblems, blood glucose, #bloodglucose, blood sugar, #bloodsugar, #diabeticstruggles, #lifewithdiabetes, #type1diabetes, #type2diabetes, #insulin4all, #thisisdiabetes, #stopdiabetes, #fingerprick	Antigua and Barbuda, Australia, The Bahamas, Barbados, Belize, Canada, Dominica, Grenada, Guyana, Ireland, Jamaica, Malta, New Zealand, St Kitts and Nevis, St Lucia, St Vincent and the Grenadines, Trinidad and Tobago, United Kingdom, United States of America
Filipino	Diyabetis, #diyabetis, #dyabetiko, Dugo glucose, #asukalsadugo, problema sa diabetes	Philippines

French	insuline, #insuline, Diabète, #Diabète, diabétique, #diabétique, #problèmedediabétique, glucose sanguin, #glucosesanguin, #maviedediabétique, #diabètetype1, #diabète de type 2, #diabètedetype1, #diabetetype2, #mondiabete	France, Canada, Belgium, Switzerland, Congo-Kinshasa, Congo-Brazzaville, Côte d'Ivoire, Madagascar, Cameroon, Burkina Faso, Niger, Mali, Senegal, Haiti, Benin
German	Diabetiker, #Diabetiker, #diabetikerproblem, Blutzucker, #blutzucker, #meindiabetes, #Diabetikerleben, #diabetestyp1, #Typ1Diabetes, #diabetestyp2, #Typ2Diabetes	Germany, Belgium, Austria, Switzerland, Luxembourg, Liechtenstein
Greek	ινσουλίνη, #ινσουλίνη, Διαβήτης, #Διαβήτης, διαβητικός, #διαβητικός, γλυκόζη αίματος, #γλυκόζηςστοαίμα, ζωή με διαβήτη, #διαβήτηςτύπου1, #διαβήτης τύπου 2	Greece, Cyprus
Hausa	ciwon diabet, #ciwonsukari, jini glucose, #jiniglucose	Nigeria, Niger, Cameroon, Chad, Sudan
Hindi	इंसुलिन, #इंसुलिन, मधुमेह, #मधुमेह, मधुमेह की समस्या, रक्त द्राक्ष - शर्करा, मधुमेह के संघर्ष, के साथ जीवन मधुमेह, टाइप 1 मधुमेह, टाइप 2 मधुमेह, madhumeh	India
Indonesian	gula darah, #guladarah, perjuangan diabetes, hidup dengan diabetes, diabetes tipe 1, #diabetestipe1, diabetes tipe 2, #diabetestipe2	Indonesia
Italian	glucosio nel sangue, #glucosionelsangue, diabete di tipo 1, #diabeteditipo1, diabete di tipo 2, #diabeteditipo2	Italy, San Marino, Switzerland, Vatican City
Japanese	インスリン, #インシュリン, 糖尿病, #糖尿病, 糖尿病, #糖尿病の, 糖尿病の問題, 血糖, #血糖, 1型糖尿病, #1型糖尿病, 2型糖尿病, #2型糖尿病, #私の糖尿病	Japan
Korean	인슐린, #인슐린당뇨병, #당뇨병, 당뇨병 환자, #당뇨병 환자, 혈당, #혈당, 당뇨병 투쟁, 제 1 형 당뇨병, 제 2 형 당뇨병	Korea
Malay	pesakit kencing manis, #pesakitkencingmanis, glukosa darah, #glukosadarah, diabetes jenis 1, diabetes jenis 2	Malaysia, Brunei, Indonesia, Singapore, The Philippines, Thailand
Norwegian	blodsukker, #blodsukker, #mindiabetes	Norway

Polish	cukrzyca, #cukrzyca, cukrzycowy, #cukrzycowy, glukoza we krwi, #glukozawekrwi, #cukrzycatypu1, #cukrzycatypu2, #mojacukrzyca, życie z cukrzycą	Poland
Portuguese	#problemasdiabéticos, glicose no sangue, #glicosenosangue, #vidacomdiabetes	Brazil, Mozambique, Angola, Portugal, Guinea-Bissau, East Timor, Equatorial Guinea, Cape Verde, São Tomé and Príncipe
Romanian	insulină, #insulinăDiabet, #Diabet, glucoza din sange, #glucozadinsange, diabet de tip 1, diabet de tip 2	Romania, Republic of Moldova
Russian	инсулин, #инсулин, сахарный диабет, #сахарный диабет, диабетом, #диабетический, содержание глюкозы в крови, #жизньсдиабетом, диабет 1 типа, #диабет1типа, диабет 2 типа, #диабет2типа	Russia, Belarus, Kyrgyzstan, Kazakhstan
Spanish	insulina, #insulina, diabético, #diabético, #problemas de la diabetes, glucosa en sangre, #glucosaenlasangre, #Diabetestipo1, #Diabetestipo2, #detenerladiabetes, #estoesdiabetes	Argentina, Bolivia, Chile, Colombia, Costa Rica, Cuba, Dominican Republic, Ecuador, El Salvador, Equatorial Guinea, Guatemala, Honduras, Mexico, Nicaragua, Panama, Paraguay, Peru, Spain, Uruguay
Swahili	Insulini, #insulini, Kisukari, #kisukari, Glucose ya damu, kisukari type 1, kisukari type 2	Tanzania, Kenya, Uganda, The Democratic Republic of Congo, the Comoros Islands
Swedish	blodsocker, #blodsocker, #typ1diabetes, #typ2diabetes	Sweden
Thai	กลูโคส, #กลูโคส, อินซูลิน, #อินซูลิน, โรคเบาหวาน, #โรคเบาหวาน, การต่อสู้กับโรคเบาหวาน, ระดับน้ำตาลในเลือด, #ระดับน้ำตาลในเลือด, การต่อสู้กับโรคเบาหวาน, โรคเบาหวานประเภท 1, โรคเบาหวานประเภท 2, โรคเบาหวานของฉัน	Thailand, Vietnam, Laos
Turkish	ensülin, #ensülin, şeker hastalığı, #şeker hastalığı, şeker hastası, #şekerhastası, kan şekeri, #kan şekeri, tip 1 diyabet, #tip1diyabet, tip 2 diyabet, #tip2diyabet	Albania, Azerbaijan, Bosnia and Herzegovina, Bulgaria, Greece, Northern Cyprus, Kosovo, the Republic of Macedonia, Moldova, Montenegro, Romania, Russia, Serbia, Syria, Turkey, Turkmenistan,Uzbekistan
Urdu	انسولین, #انسولین, ذیابیطس, #ذیابیطس, خون میں گلوکوز, #خون میں گلوکوز, ٹائپ 1 ذیابیطس, ٹائپ 2 ذیابیطس,	Pakistan, India, Afghanistan, Saudi Arabia
Vietnamese	Bệnh tiểu đường, #Bệnhtiểuđường, Bệnh tiểu đường., #mắcbệnhđáiđường, đường huyết,	Vietnam

	#đường huyết, bệnh tiểu đường loại 1, bệnh tiểu đường loại 2	

Appendix 4: Details on data processing for the analysis of diabetes-related tweets

We followed most of the same processes described by *Ahne et al.* in their Supplementary online material 2.[1]

Data collection

We accessed the API Standard v1.1 by applying for a Twitter Developer account.[2] We connected to Twitter's API using the Python library Tweepy and streamed the keywords to collect tweets together with users' metadata (attributes provided on user's profile, such as user screen name, user location, user description.).[3]

Data representation

We used FastText implementation in the Gensim package to process each word into a vector to extract meaningful semantic relationships.[4] The average of each tweet's word vector representations was then used to model it and similarities in their semantics were analyzed.

Geolocation process

We aimed to keep only geolocated tweets for this study. To do so, we first looked at users' locations, provided by users themselves in their public profile, and removed all tweets without such location. We then grouped all locations and manually detected which ones were fake in order to exclude it (e.g. "the Internet", "in Hell"). We also replaced all contractions to full words (eg. "CA, USA" to "California, United States of America") to facilitate geolocation by the Python package geograpy4.[5] We then applied the package to the different grouped locations in order to access the latitude and the longitude of the location associated with each tweet. If the determined place was a country, latitude and longitude were identified in the center of the country. We evaluated the overall precision of this geolocation step as 85%.

Tweets translation

In order to apply unique classifiers to all tweets, we translated all tweets originally not written in English to English using the Python package deep-translator, a free and unlimited python API for Google Translate.[6]

Personal content classifier / Jokes classifier

We developed two classifiers that aimed to filter out institutional tweets and jokes in order to keep only tweets with personal content and not jokes or irony about diabetes. A tweet was considered as personal if the user expressed his feelings and own experiences, dealing with his own diabetes or a relative one. A tweet was considered as a joke/irony if diabetes was used as an insult or with a sugar-related joke. To train these two classifiers, three authors (AA, CB, GF) manually labelled 1648 randomly chosen tweets for the personal classifier and 1398 for the jokes one. We used *Bidirectional Encoder Representations from Transformers* (BERT), a machine learning technique for natural language processing pre-training developed by Google.[7] More precisely, we applied *BERTweet*, a pre-trained model for English Tweets.[8] The overall accuracy of the personal content classifier was 88% and the accuracy from the jokes classifier was 94%.

Gender and Type of diabetes classifier

Following the scripts by *Ahne et al.*, we trained classifiers to predict gender (male, female, unknown) and type of diabetes (type 1, type 2, unknown) from each user.[88] Three authors (AA, CB, GF) manually labelled 1670 tweets from different regions of the world to better match our dataset. A SVM was trained and reached an accuracy of 86% for the gender classifier and 74% for the type of diabetes classifier.

Topic extraction

All tweets are represented via their word vector representations. Then, a K-means algorithm was applied to the tweets in each region. To define the optimal number of clusters *k*, the silhouette score average for *k* between 4 and 24 was calculated and silhouette analysis applied.[9]

Emotion classifier

Initially, two authors (CB, GF) labelled 1000 randomly chosen tweets according to the main emotion in the tweet text. In order to increase the accuracy of our classifier, we combined our dataset with online labelled datasets of emotions which led to the extension of 47,000 tweets from Github and Kaggle.[10–12] We then trained a Calibrated Linear Support Vector classifier to predict the probability of a tweet belonging to each of the four emotions.[13] We applied this classifier to all tweets to predict the probability of a tweet belonging to each of the four emotions.

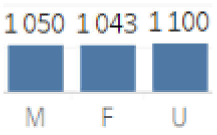
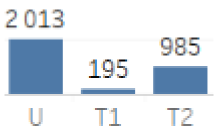
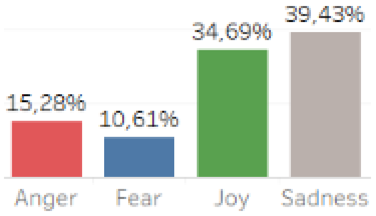
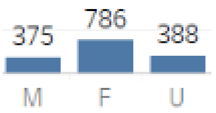
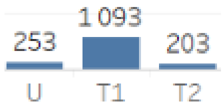
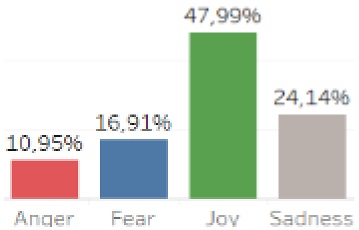
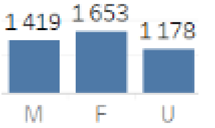
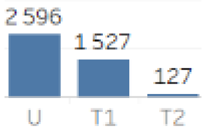
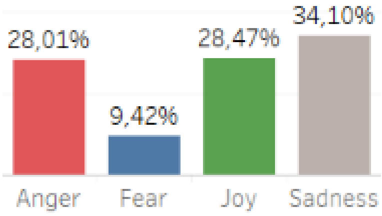
This led to the creation of a dataset of 47,926 texts labelled as one of the following emotions: joy, anger, fear or sadness. We based our classifier on the following scripts: https://thecleverprogrammer.com/2021/02/19/text-emotions-detection-with-machine-learni ng/. We calibrated the classifier in order to predict the probability of a tweet to belong to each of the 4 classes.[14] The calibrated SVM was trained and reached an accuracy of 80.56% (precision to predict joy only: 88%, precision to predict fear only: 94%, precision to predict anger only: 93%, precision to predict sadness only: 89%).

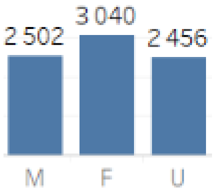
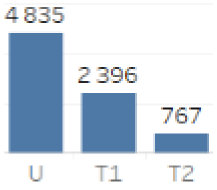
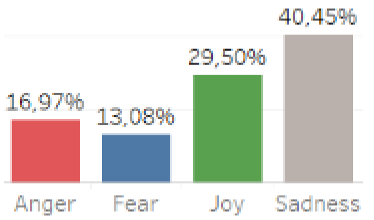
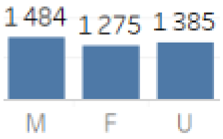
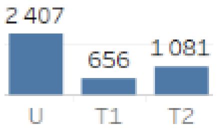
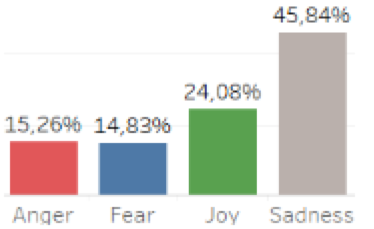
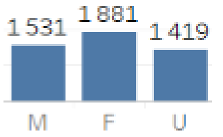
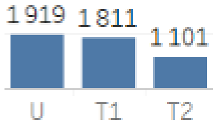
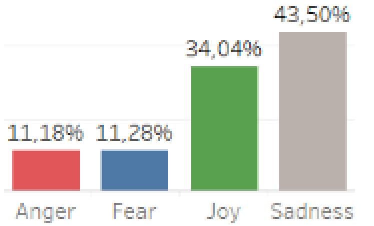
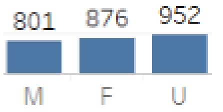
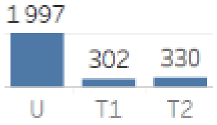
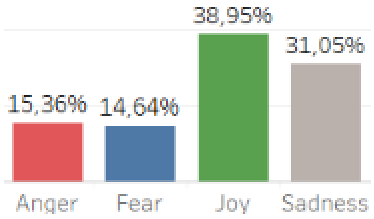
Details

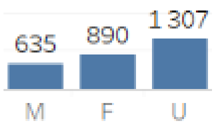
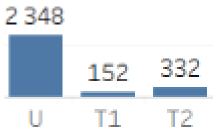
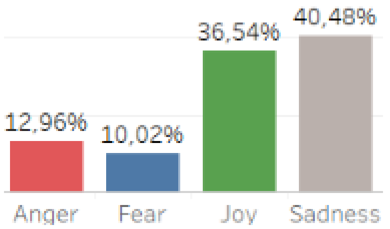
Python (V.3.6) with the packages scikit-learn (machine learning algorithms and data preprocessing methods), gensim (text processing, word representation), and statsmodels (module for the estimation of many different statistical models) were exploited for the analysis.[15–17] Tableau (2020.4) and OpenStreetMap 2021 were used to visualize the data.

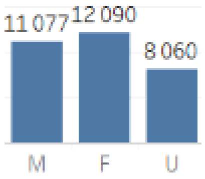
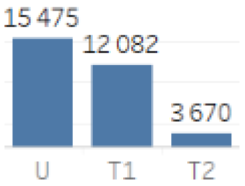
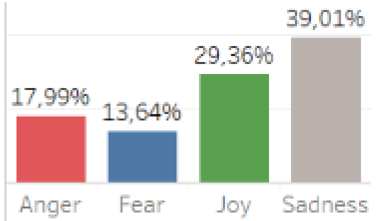
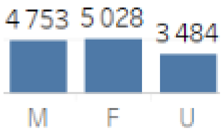
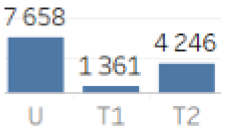
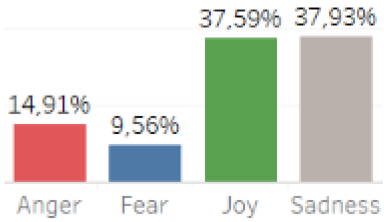
Additional references

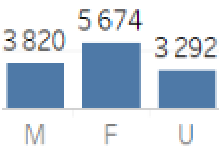
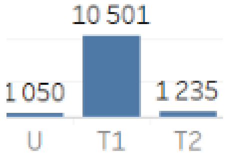
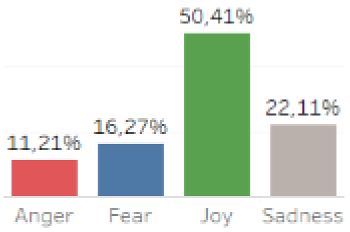
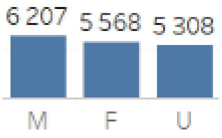
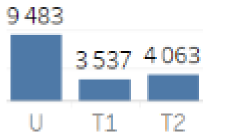
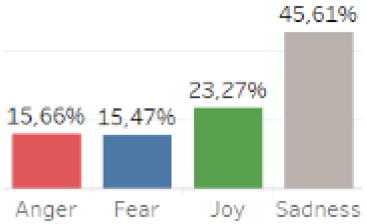
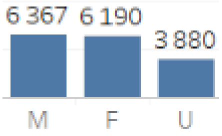
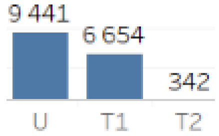
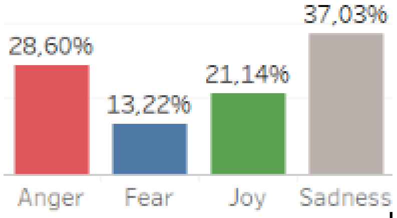
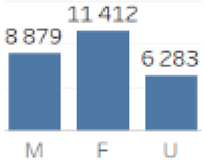
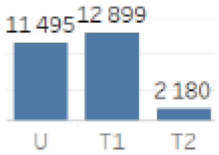
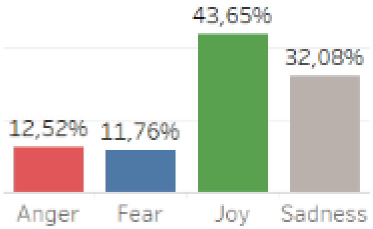
Appendix 5: Overview of the topics of interest by region

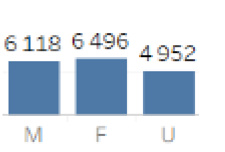
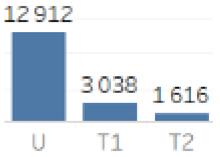
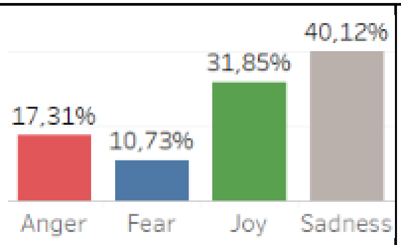
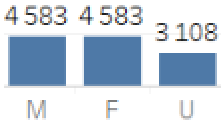
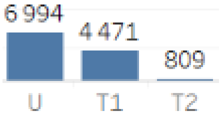
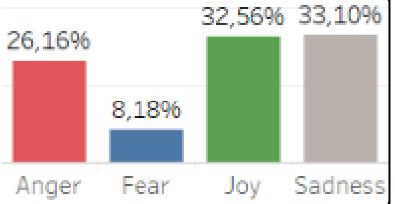
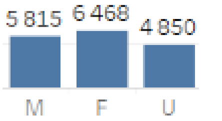
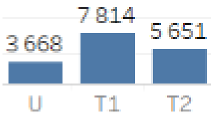
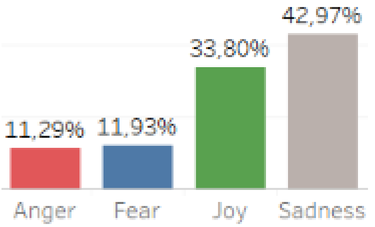
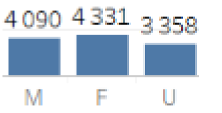
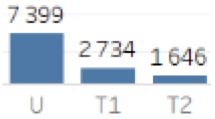
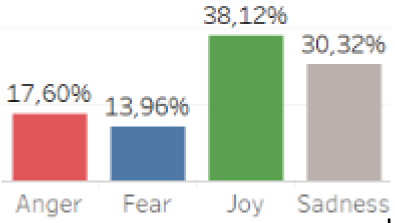
East Asia and Pacific							
Cluster n°	**Topic label**	**Topwords**	**Gender**	**Diabetes type**	**Tweet description**	**Average probability of emotions**	**Number of tweets in topic**
0	Food restrictions/issues implied by diabetes	sugar, eat, glucose, insulin, like, drink, eating			Food habits as diabetic and implications on blood glucose levels		3193 (N=10.2%)
1	Type 1 diabetes communities	#t1d, #diabetes, one, #dsma, #gbdoc, #bloodsugar, #type1diabetes			Tweets from t1, dsma and gbdoc community sharing daily experiences		1549 (N=4.9%)
2	Insulin affordability	insulin, one, need, people, would, get, like			Tweets about luck to be able to afford insulin, comments about prices in the US		4250 (N=13.5%)

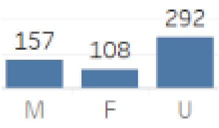
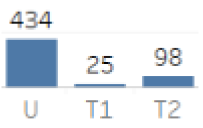
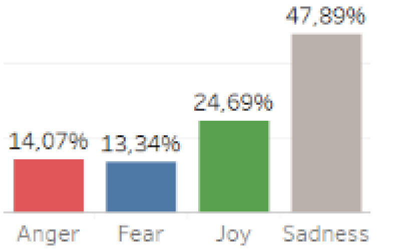
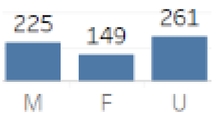
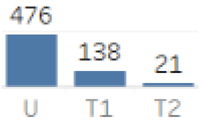
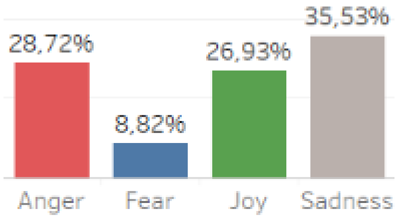
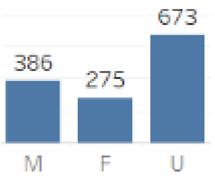
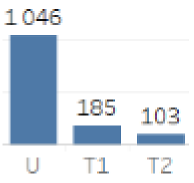
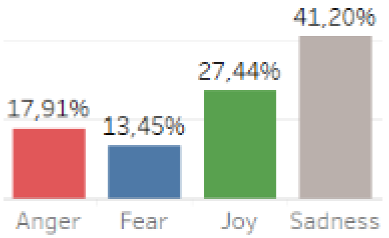
3	Explaining what it is like to have diabetes	one, get, would, like, people, know, type			Comments and reactions about what poeple think about diabetes compared to what it is like		7998 (N=25.5%)
4	Diabetes related complications and family history	heart, type, one, two, disease, high, insulin			Explaining what are the potential complications of diabetes. Comments about family history of diseases		4144 (N=13.2%)
5	Life and experiences since diagnosis	one, two, type, years, insulin, got, day			Explaining habits with family, dealing with their diabetes or own diabetes since diagnosis		4831 (N=15.4%)
6	Glycemic control	glucose, blood, sugar, test, insulin, levels, need			Tweets about tracking blood glucose levels, giving levels examples		2629 (N=8.4%)

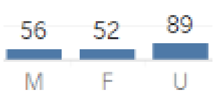
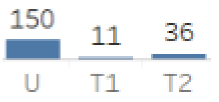
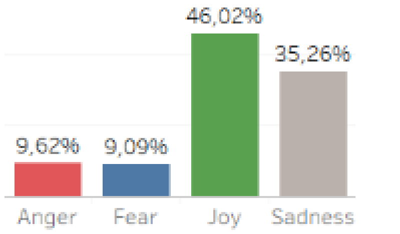
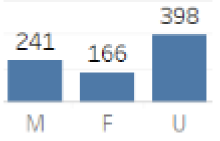
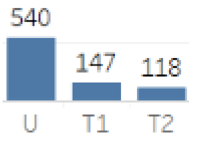
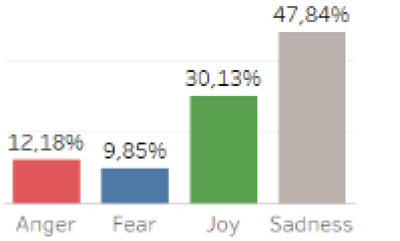
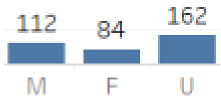
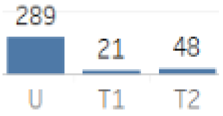
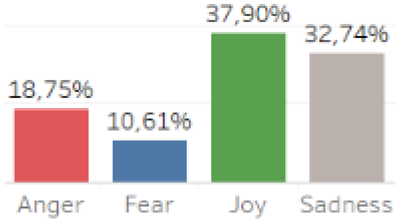
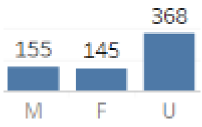
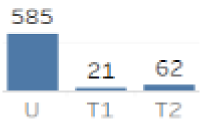
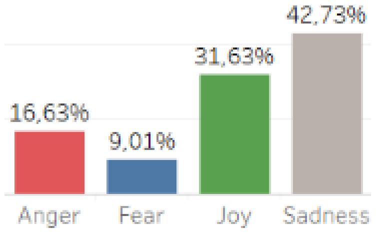
7	Reactions to jokes and news about diabetes	sweet, got, get, like, one, two, insulin			Angry comments about jokes, irony and cliches about diabetes		2832 (N=9.0%)

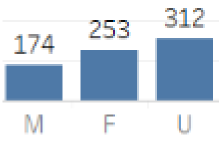
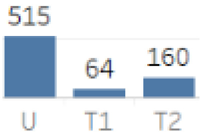
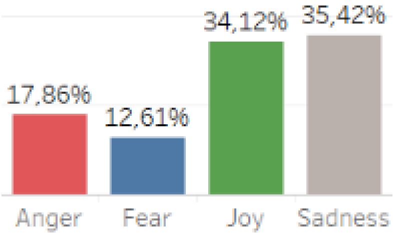
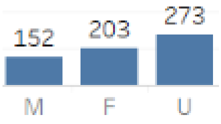
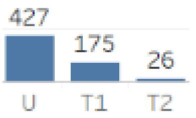
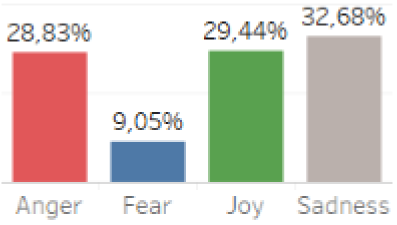
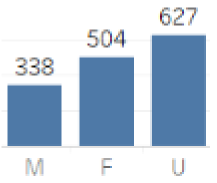
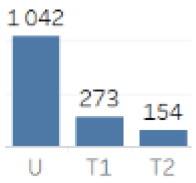
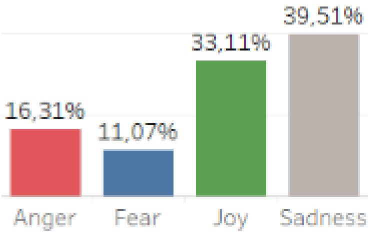
Europe and Central Asia							
Cluster n°	**Topic label**	**Topwords**	**Gender**	**Diabetes type**	**Tweet description**	**Average of emotions**	**Number of tweets in topic**
0	Living with diabetes	one, would, people, type, know, like, get			Sharing information, advices and experiences about life with diabetes		31227 (N=17.7%)
1	Food habits	sugar, eat, one, insulin, glucose, eating, carbs			Impact of food and specific diets on blood glucose levels, share of food advices and experiences		13265 (N=7.5%)

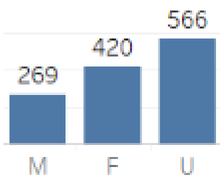
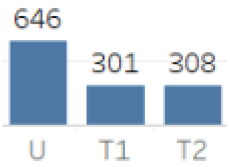
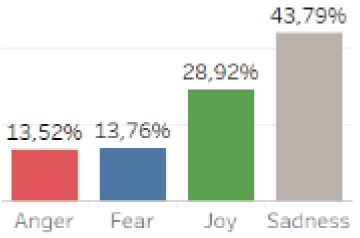
2	Type 1 diabetes communities	#gbdoc, #t1d, #diabetes, #type1diabetes, one, day, get			Tweets with #gbdoc and #t1d hashtags, motivation and advices		12786 (N=7.3%)
3	Diabetes related complications and family history	type, one, heart, two, people, disease, cancer			Testimony about family complications and events		17083 (N=9.7%)
4	Insulin access	insulin, one, would, people, cannot, get, need			Reactions to insulin affordability in the USA, gratitude to be able to access insulin		16437 (N=9.3%)
5	Daily management of diabetes	one, insulin, get, day, today, two, time			Tweets about daily management of diabetes (medical appointments, dealing with blood glucose…)		26574 (N=15.1%)

<table>
<tr><td>6</td><td>Diabetes-related stories</td><td>like, got, one, get, insulin, type, oh</td><td>
</td><td>
</td><td>Sharing stories and reacting to others stories about diabetes</td><td>
</td><td>17566 (N=10%)</td></tr>
<tr><td>7</td><td>Insulin and insulin supplies</td><td>insulin, pump, get, day, today, two, time</td><td>
</td><td>
</td><td>Testing insulin supplies, sharing past experiences and mistakes about insulin</td><td>
</td><td>12274 (N=7%)</td></tr>
<tr><td>8</td><td>Life changes since diagnosis</td><td>type, one, two, years, diagnosed, year, old</td><td>
</td><td>
</td><td>Sharing what happened, new habits and what changed since diagnosis</td><td>
</td><td>17133 (N=9.7%)</td></tr>
<tr><td>9</td><td>Glycemic control and tests</td><td>blood, glucose, sugar, levels, test, insulin, high</td><td>
</td><td>
</td><td>Sharing experiences with blood glucose tests and advices to lower it</td><td>
</td><td>11779 (N=6.7%)</td></tr>
</table>

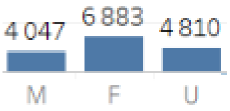
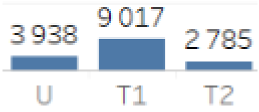
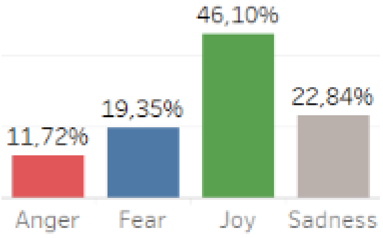
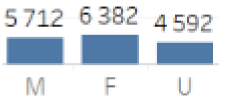
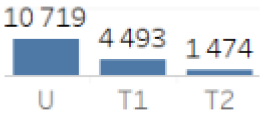
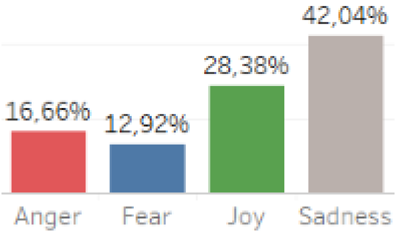
Latin America and the Caribbean							
Cluster n°	Topic label	Topwords	Gender	Diabetes type	Tweet description	Average of emotions	Number of tweets in topic
0	Complications and comorbidities	high, hypertension, cancer, pressure, blood, heart, people			Sharing diagnosis of a complication, potential risks of comorbidities		557 (N=12.2%)
1	Insulin issues	insulin, need, one, people, cannot, get, like			Tweets about insulin pricing, affordability, resistance..		635 (N=13.9%)
2	Life with diabetes	people, one, get, know, like, insulin, would			Explaining what is diabetes to uninformed people and real issues		1334 (N=29.3%)

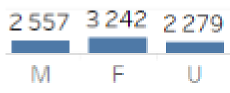
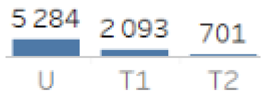
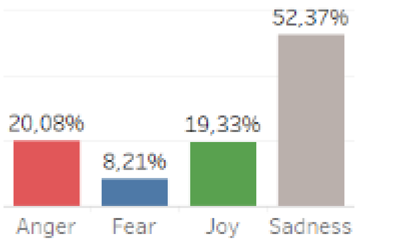
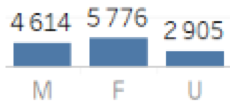
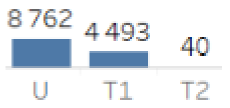
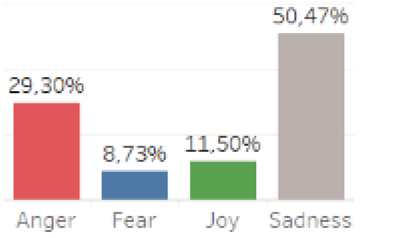
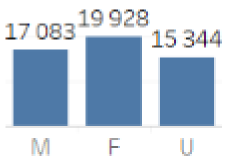
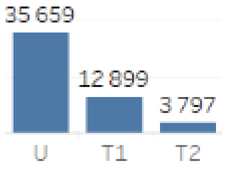
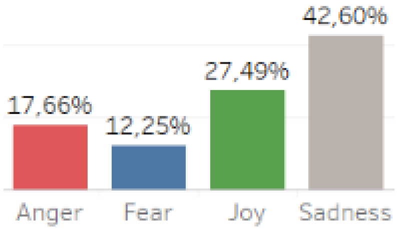
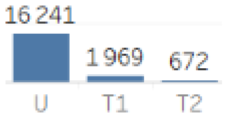
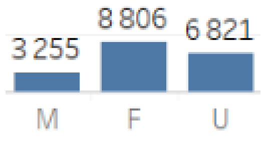
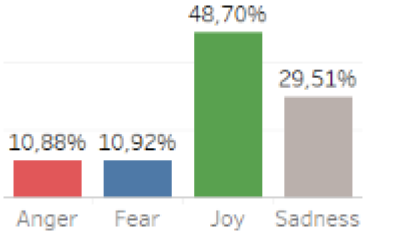
3	Love and support	sweet, much, love, cute, hope, nick, #bloodsugar			Sharing love and support messages including to popstar Nick Jonas		197 (N=4.3%)
4	Experiences from relatives living with diabetes	two, years, one, old, ago, day, mother			Sharing relatives experiences with diabetes, asking for help to assist family		805 (N=17.7%)
5	Glycemic control	glucose, sugar, blood, test, insulin, levels, eat			Asking and sharing advices to deal with and lower blood glucose levels		358 (N=7.9%)
6	Reactions to jokes about diabetes	like, sugar, one, get, eat, insulin, two			Reactions and answers to jokes and clichés about diabetes		668 (N=14.7%)

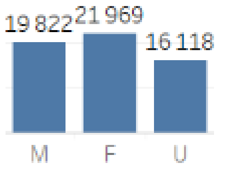
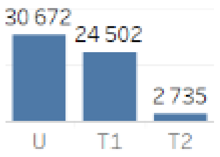
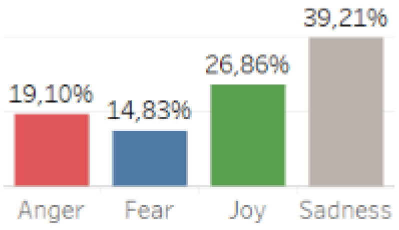
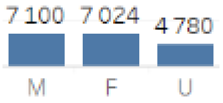
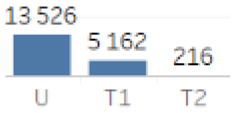
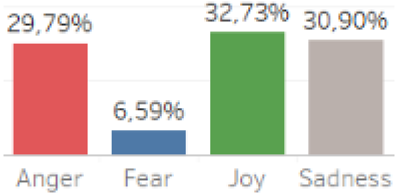
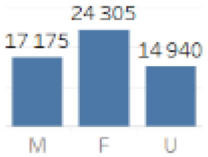
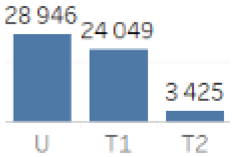
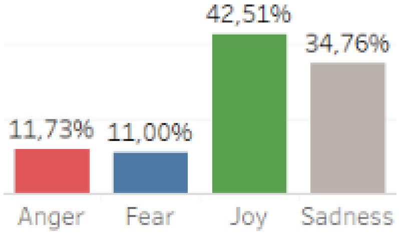
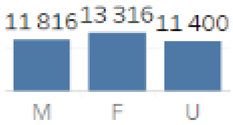
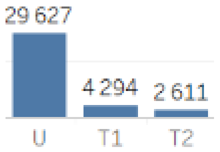
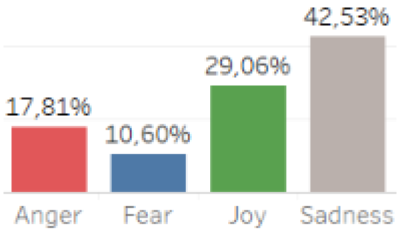
Middle East and North Africa							
Cluster n°	**Topic label**	**Topwords**	**Gender**	**Diabetes type**	**Tweet description**	**Average of emotions**	**Number of tweets in topic**
0	Food and glycemic control	glucose, blood, sugar, insulin, high, eat, low			Influence of food on blood glucose levels, diet examples		739 (N=18.1%)
1	Insulin and insulin supplies	insulin, one, need, take, pump, cannot, two			Sharing experiences and advices about the use of insulin and insulin supplies		628 (N=15.4%)
2	Daily life and habits	one, like, get, people, would, know, got			Tweets of users sharing their daily life and struggles to manage diabetes		1469 (N=35.9%)

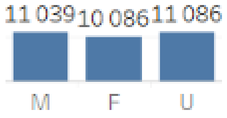
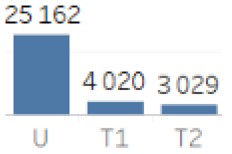
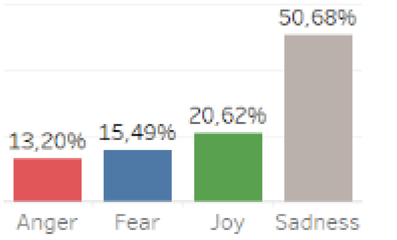
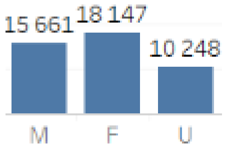
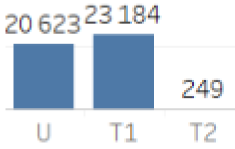
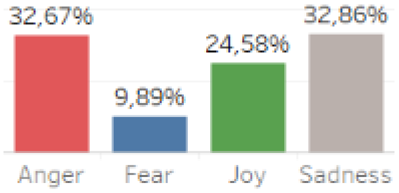
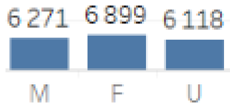
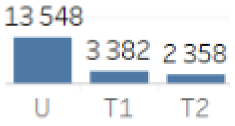
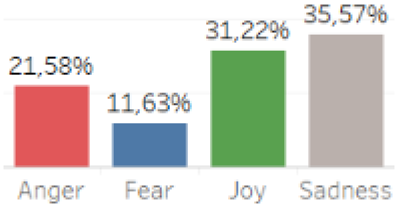
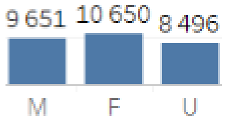
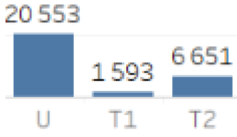
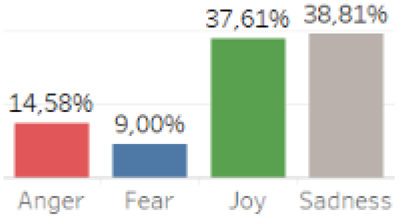
3	Experiences from relatives living with diabetes	one, type, two, years, insulin, people, old			Sharing what is it like to live around diabetes, age of relatives with diabetes, time since diagnosis		1255 (N=30.7%)

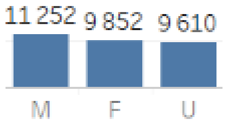
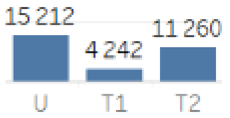
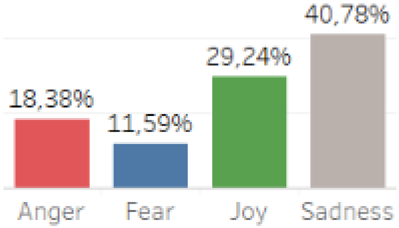
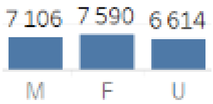
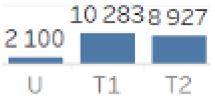
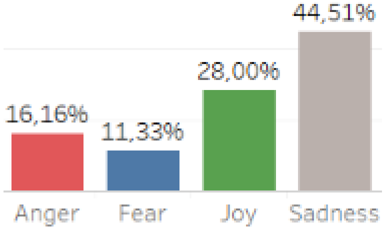
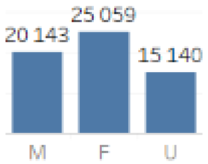
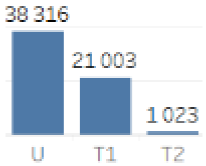
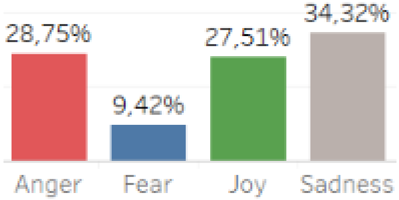
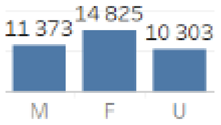
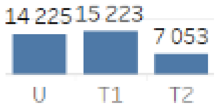
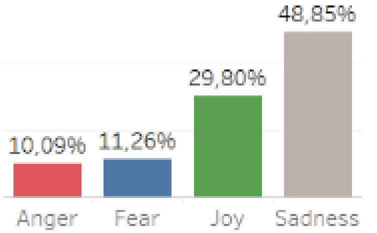
North America

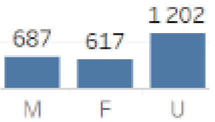
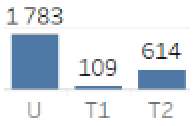
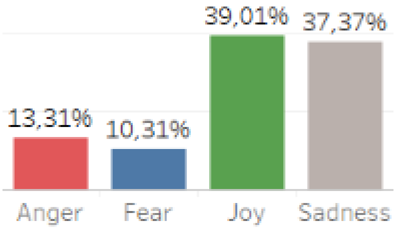
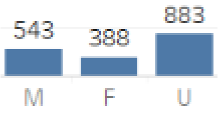
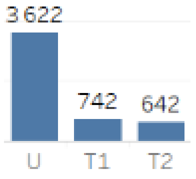
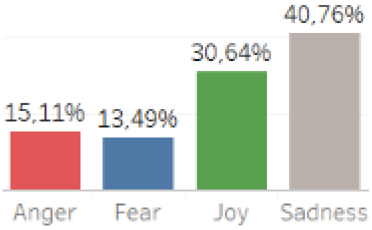
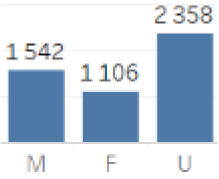
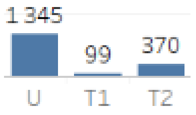
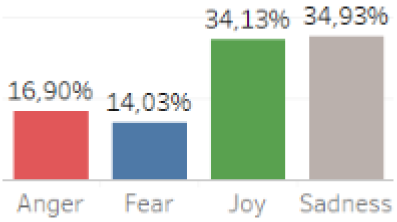
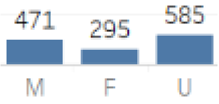
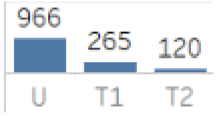
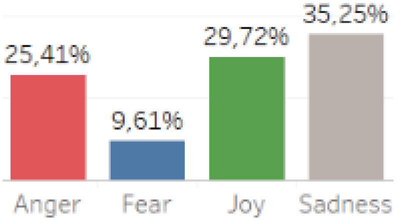
Cluster n°	Topic label	Topwords	Gender	Diabetes type	Tweet description	Average of emotions	Number of tweets in topic
0	Type 1 diabetes communities	#t1d, #diabetes, #bloodsugar, #type1diabetes, one, #diabetic, #dsma			Tweets with Type 1 and dsma hashtags		15740 (N=2.8%)
1	Explaining insulin struggles to uninformed people	would, could, insulin, one, like, get, know			Stories about relatives or own struggles of living with diabetes (insulin, insurance..)		16686 (N=2.9%)

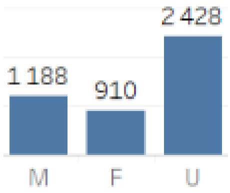
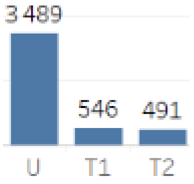
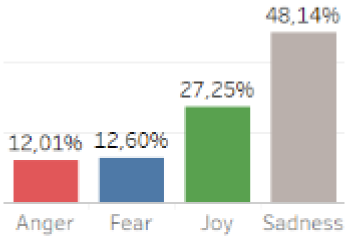
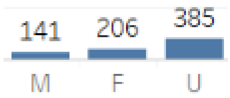
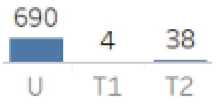
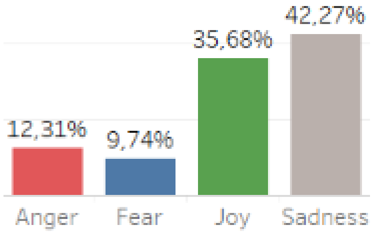
2	Inability to afford insulin	cannot, get, insulin, one, eat, like, wait			Tweets about insulin affordability, struggles of not being able to afford insulin		8078 (N=1.4%)
3	Consequences of insulin unaffordability	insulin, afford, cannot, people, dying, die, could			Tweets about potential complications due to insulin unaffordability, fight to get insurance, insulin rationing		13295 (N=2.3%)
4	Explaining risks of living with diabetes	like, know, people, get, one, really, think			Reactions to comments of unaware people about diabetes and explaining risks of living with diabetes		52355 (N=9.2%)
5	Glucose tests	glucose, test, hour, drink, three, one, blood			Tweets about glucose tests and pregnancy		18882 (N=3.3%)

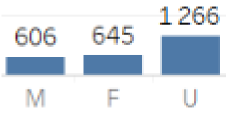
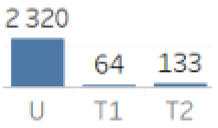
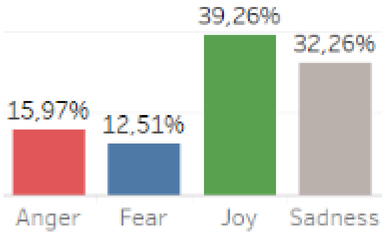
6	Healthcare issues	one, people, insulin, get, type, health, care			Tweets about political choices (election, reforms), comparing healthcare in the USA and Canada		57909 (N=10.2%)
7	Insulin prices increase	insulin, pump, need, hundred, one, expensive, like			Explaining issues of constant insulin prices increase and overpricing		18904 (N=3.3%)
8	Sharing daily life	one, get, day, insulin, got, two, today			Tweets about what happened in the day		56420 (N=9.9%)
9	Jokes about diabetes	got, like, get, one, as, mum, shit, dad			Jokes about people with diabetes and reactions to such jokes		36532 (N=6.4%)

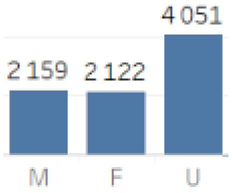
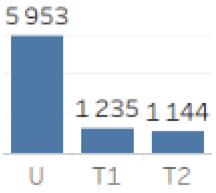
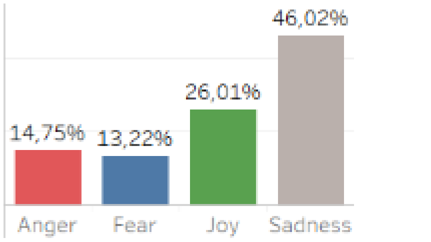
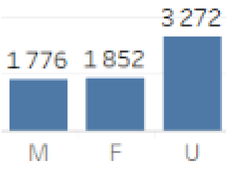
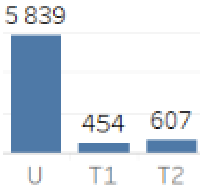
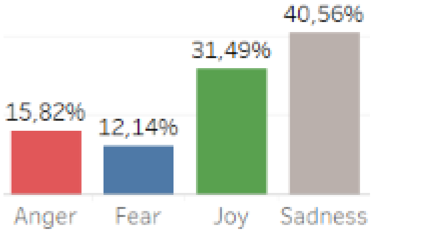
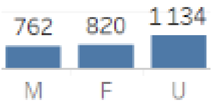
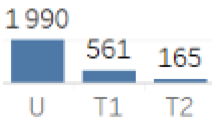
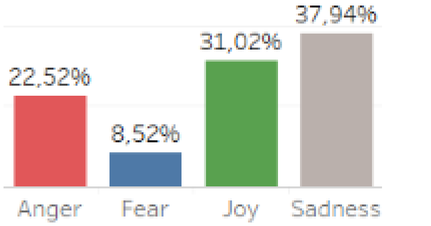
10	Comorbidities	heart, disease, cancer, people, one, like, get			Family medical history and comorbidities		32211 (N=5.7%)
11	Insulin pricing including insurance	insulin, hundred, one, insurance, cost, price, pay			Giving prices of insulin, money spent on insulin with and without insurance		44056 (N=7.8%)
12	Glycemic control and complications	blood, sugar, high, pressure, glucose, insulin, good			Blood sugar levels, blood pressure,managing blood sugar levels		19288 (N=3.4%)
13	Food restrictions	sugar, eat, like, drink, one, glucose, eating			Restrictions and cut out foods, difficulty of such restrictions		28797 (N=5.1%)

14	Food influence on blood sugar	insulin, glucose, sugar, diet, weight, two, type			New food habits and diet and impact of food blood glucose levels		30714 (N=5.4%)
15	Confusion type 1 and type 2 diabetes	type, one, two, diagnosed, get, people, know			Explaining differences between type 1 and type 2, misdiagnosis		21310 (N=3.8%)
16	Costs related to diabetes management	insulin, like, get, need, one, people, pump			Describing all costs implied by supplies: insulin, insulin pump, CGM, pen		60342 (N=10.6%)
17	Life since diagnosis	years, two, one, old, year, type, ago			Own and relatives diagnosis and what changed in life since diagnosis		36501 (N=6.4%)

South Asia							
Cluster n°	Topic label	Topwords	Gender	Diabetes type	Tweet description	Average of emotions	Number of tweets in topic
0	Food habits	eat, sugar, sweet, #letscheatdiabetes, food, glucose, like			Tweets about food restrictions and habits, sharing advices about food		2506 (N=15.7%)
1	Advices about glycemic control	glucose, blood, sugar, high, levels, insulin, level			Sharing experiences and giving advices to deal with blood glucose levels		5006 (N=31.4%)
2	Anecdotes about life with diabetes	one, people, like, get, know, also, two			Life since diagnosis, dealing with doctors, relatives diabetes		1814 (N=11.4%)
3	Insulin use	insulin, one, hundred, take, two, need, get			Describing used insulin supplies and habits		1351 (N=8.5%)

4	Reaching celebrities for support	patient, one, sir, father, and, hundred, help			Tagging "famous" users and asking for their help		4526 (N=28.4%)
5	Glucose and tests	glucose, sweet, patient, reconnaissance			Sharing relatives experiences as patients with diabetes		732 (N=4.6%)

Sub Saharan Africa							
Cluster n°	**Topic label**	**Topwords**	**Gender**	**Diabetes type**	**Tweet description**	**Average of emotions**	**Number of tweets in topic**
0	Glucose guardian	glucose, sugar, guardian, need, blood, daddy, lady			Searching and asking for a glucose guardian to pay for insulin, thanking glucose guardians		2517 (N=12.3%)

1	Experiences from relatives living with diabetes	one, people, two, blood, know, high, like			Tweets about experiences and struggles of dealing with own or relatives diabetes		8332 (N=40.7%)
2	Food influence on blood sugar levels	sugar, glucose, get, one, like, please, go			Diet and food restrictions, new food habits and diet, implications on blood sugar levels		6900 (N=33.7%)
3	Insulin	insulin, buy, need, drugs, help, hundred, get			Asking for help and sharing advices about insulin, reactions to insulin stories		2716 (N=13.3%)

Appendix 6: Concepts, keywords and thresholds regarding diabetes for a VDCS using ALTRUIST

Concepts	Keywords	Threshold
Diabetes	diabetes	0.33
Diabetes devices and supply	pump, strips, monitor, lancet, pen, syringe, needle, swab, gauze, cgm, continuous glucose monitor, Care Touch, Freestyle Libre, Dexcom, Eversence, Rite aid, truemetrix	0.26
Diagnosis	diagnosis	0.40
Food	food, nutrition, diet	0.27
Symptom	symptom, sweating, fatigue, dizziness, dizzy, hungry, hypoglycaemia, hyperglycaemia, shaky, trembling, palpitations, irritated, tearful, moody, pale, thirst	0.38
Family	family	0.27
Weight	weight, overweight, obesity	0.40
Sleep	sleep	0.36
Mental health	depression, distress, burnout, anxiety, exhaustion, anxious, burden	0.41
Sport	exercising	0.32
Politic	politic	0.26
Death	death	0.40
Affordability	affordability, money, insulin access	0.53
Medication	medication, insulin, metformin, fortamet, humulin, novolin, novolog, flexpen, fiasp, precose, glyset	0.29
Glycemic control	hba1c	0.38
Healthcare system	healthcare system, insurance, social security	0.33
Comorbidities	hypertension, cancer, asthma, dyslipidemia, hypothyroidism, hyperlipidemia	0.37

Complications	heart disease, osteoarthrosis, amputation, retinopathy, nerve damage, foot problem, heart attack, stroke, nephropathy, kidney problem, neuropathy, kidney problem, neuropathy, gum disease, urinary infection, erectile dysfunction, impotence, ketoacidosis, dka, heart failure, vascular disease, myocardial infarction, cardiovascular disease, ischemic heart disease	0.30
Diabetes risk factors	alcohol, inactivity, smoking, polycystic ovary syndrome, high blood pressure, cholesterol, drugs	0.44
Hypoglycemia	hypoglycemia, hypoglycaemia	0.47