ALTRUIST: A Python Package to Emulate a Virtual Digital Cohort Study Using Social Media Data

Charline Bour, Abir Elbeji, Luigi De Giovanni, Adrian Ahne, and Guy Fagherazzi

***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 (SM) 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, SM offer unstructured but insightful information about patients' disease burden. Yet, there is limited guidance on analyzing SM data as a prospective cohort. We presented the concept of virtual digital cohort studies (VDCS) as an approach to replicate cohort studies using SM data. In this paper, we introduce ALTRUIST, an open-source Python package enabling standardized generation of VDCS on SM. 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 SM 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 data from various online communities of patients, complementing traditional epidemiological methods and promoting minimally disruptive health research.**

***Index Terms*—Cohort, digital health, natural language processing, python, social media.**

I. Introduction

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 [1]. The methodology of such a study can be divided into five steps: 1) recruitment of the participants out of the target population, 2) acquisition 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 [2], [3].

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 [3], [4].

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 [5]. 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 [6], [7]. Social media users leave behind them a digital footprint and historical records (timelines) that can be of great interest for research purposes. In particular, in this paper, we focus on the example of Twitter, where more than 500 million tweets were sent each day in March 2023 [8], [9]. 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 [10]. 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 [11], [12].

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 [13]. 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 utilizing 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.

Manuscript received 29 June 2023; revised 10 January 2024; accepted 22 January 2024. Date of publication 5 February 2024; date of current version 11 July 2024. This work was supported by the MSDAvenir Foundation (World Diabetes Distress Study), and in part by the Luxembourg Institute of Health. Recommended for acceptance by G. Yang. *(Corresponding author: Guy Fagherazzi.)*

Charline Bour and Abir Elbeji are with the Deep Digital Phenotyping Research Unit, Department of Precision Health, Luxembourg Institute of Health, 1445 Strassen, Luxembourg, and also with the University of Luxembourg, 4365 Esch-sur-Alzette, Luxembourg (e-mail: charline.bour@lih.lu; abir.elbeji@lih.lu).

Luigi De Giovanni is with Data Integration and Analysis Unit, Luxembourg Institute of Health, 1445 Strassen, Luxembourg (e-mail: Luigi.DeGiovanni@lih.lu).

Adrian Ahne is with Akuity Care, 75017 Paris, France (e-mail: adrian.ahne@protonmail.com).

Guy Fagherazzi is with Deep Digital Phenotyping Research Unit, Department of Precision Health, Luxembourg Institute of Health, L-1445 Strassen, Luxembourg, and also with the Director of the Department of Precision Health and a Group Leader, Luxembourg Institute of Health, L-1445 Strassen, Luxembourg (e-mail: guy.fagherazzi@lih.lu).

This article has supplementary downloadable material available at https://doi.org/10.1109/TBDATA.2024.3362193, provided by the authors.

Digital Object Identifier 10.1109/TBDATA.2024.3362193

Relying on existing online data such as Twitter could circumvent some of the above-mentioned limitations (recruitment, cost, duration, research burden) 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.

This paper introduces a methodology and a Python package to generate and analyze cohort-like data from social media. It was designed to complement traditional cohort studies. By leveraging the wealth of social media data, we provide a means to enhance these studies, offering patient-centric insights that are often difficult to capture in clinical cohort settings. VDCS addresses the limitations of time and cost in traditional research and is especially valuable to include diverse health perspectives, particularly from underrepresented groups. By integrating our digital data analysis with existing cohort study frameworks, we present a synergistic approach that promises to enrich epidemiological research, providing a more complete understanding of health trends and patient experiences

II. Methods

ALTRUIST stands for virtu**AL** digi**T**al coho**R**t st**U**dy us**I**ng **S**ocial media da**T**a. It is a Python package that aims to emulate a cohort on social media data. ALTRUIST was implemented and tested in Python 3.8 [14], 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 currently provides scripts to 1) collect data using a user's login information to the Twitter API for now (more social media APIs will be added later on), 2) separate personal and non-personal information to identify users with the outcome of interest, 3) collect specific users' timelines (i.e. history data published by a user) and preprocess these timelines. 4) apply cosine similarities between tweets, outcomes, and events using all-mpnet-base-v2, a Sentence-transformer pre-trained model that generates vector representations for sentences [15], 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. However, some steps are optional and can be skipped if the package's user has already collected data elsewhere. The main steps are described in Fig. 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.

A. Data Collection

A. Original data collection: The first step is to access and collect the data. In the case of Twitter, an API [8] is currently available, but it is important to note that API functionalities and access may change in the future. For now, collected data will be stored on MongoDB, a "powerful and scalable data storage" [16]. 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 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 [17], and particularly its version BERTweet, which is a pre-trained model for English tweets [18]. For projects in languages other than English, a user can select another pre-trained model, for instance from the Huggingface model hub which is pre-trained in a specific language. The fine-tuned classifier is then applied to tweets 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 tweets 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 [19]. 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,

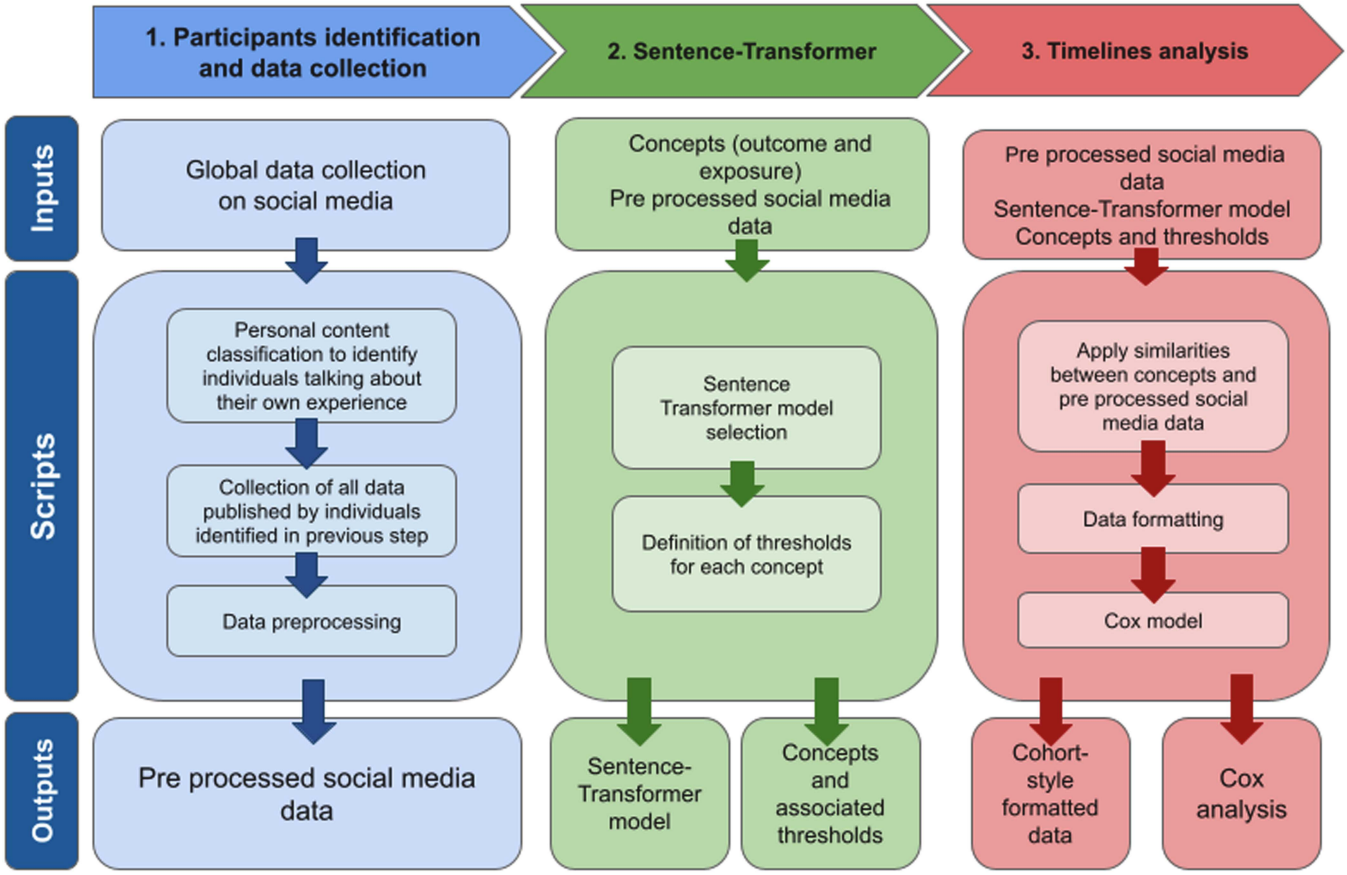

Fig. 1. ALTRUIST structure and sequential workflow.

13, 52, 51, tzinfo=datetime.timezone.utc)" [20]). 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.

B. 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.

SentenceTransformers algorithms: SentenceTransformers is a Python framework for state-of-the-art sentence, text, and image embeddings [15]. 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 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, and 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.

C. 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 [21]. 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 [22].

III. Use Case on Diabetes

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 mentioned more 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. Keywords were chosen to include common and technical terms associated with diabetes, ensuring a wide coverage of topics from general discussions to more specific medical aspects. The multilingual approach allowed us to include perspectives from users all around the world. The data collection process involved the use of Twitter's API v2 to retrieve tweets with at least one of these keywords. These keywords were streamed between 2017 and 2022 resulting in the collection of 34 million tweets. More details about the data collection, keywords, and the dataset can be found in our previous work [10].

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 I.

TABLE I
Performances of the Fine-Tuned BERTweet Personal Content Classifier on the Test Set

Accuracy	F1-Score	Precision	Recall
0.98	0.98	0.99	0.97

The 34 million diabetes-related tweets were then classified and users tweeting about their own diabetes could be identified. Fig. 2 shows the process of identifying participants for our cohort. To be included, users had to share their personal experiences with diabetes at least once. Tweets from researchers, institutions, or discussions about relatives' experiences with diabetes were categorized as "non-personal". Users identified as having diabetes had their entire timelines collected, including all tweets since account creation. Retweets were excluded. 88,057 users were identified after the classification step. 36,171 were still active at the time we started our VDCS on diabetes. Almost 60 million tweets were included for analysis after the collection and preprocessing of the timelines.

Based on previous studies [10], [22], [23], 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 1, available online: Concepts, keywords, and thresholds.

Based on previous studies [10], [23], [24], 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 1, available online. In our use case, the keywords related to "Mental health" and "Comorbidities" were defined to study their interconnection in the case of diabetes. For instance, keywords under "Mental Health" included "depression", "anxiety" and "distress" which are particularly relevant to understand the concept of diabetes distress, an emotional and psychological condition experienced by people living with diabetes [25]. Similarly, the comorbidity-related keywords provide insights into the common comorbid conditions that people with diabetes are concerned about [26]. Thus, the keywords not only help to categorize tweets but also enable us to draw meaningful insights into the complex interplay between diabetes, mental health, and physical comorbidities. They reveal patterns and themes in the social media data that are indicative of the real-life experiences and challenges faced by people living with diabetes.

The next step was to identify the best model for our analysis. Table II 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 pre-trained 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 [27]. We computed similarities between 200.000 preprocessed tweets and

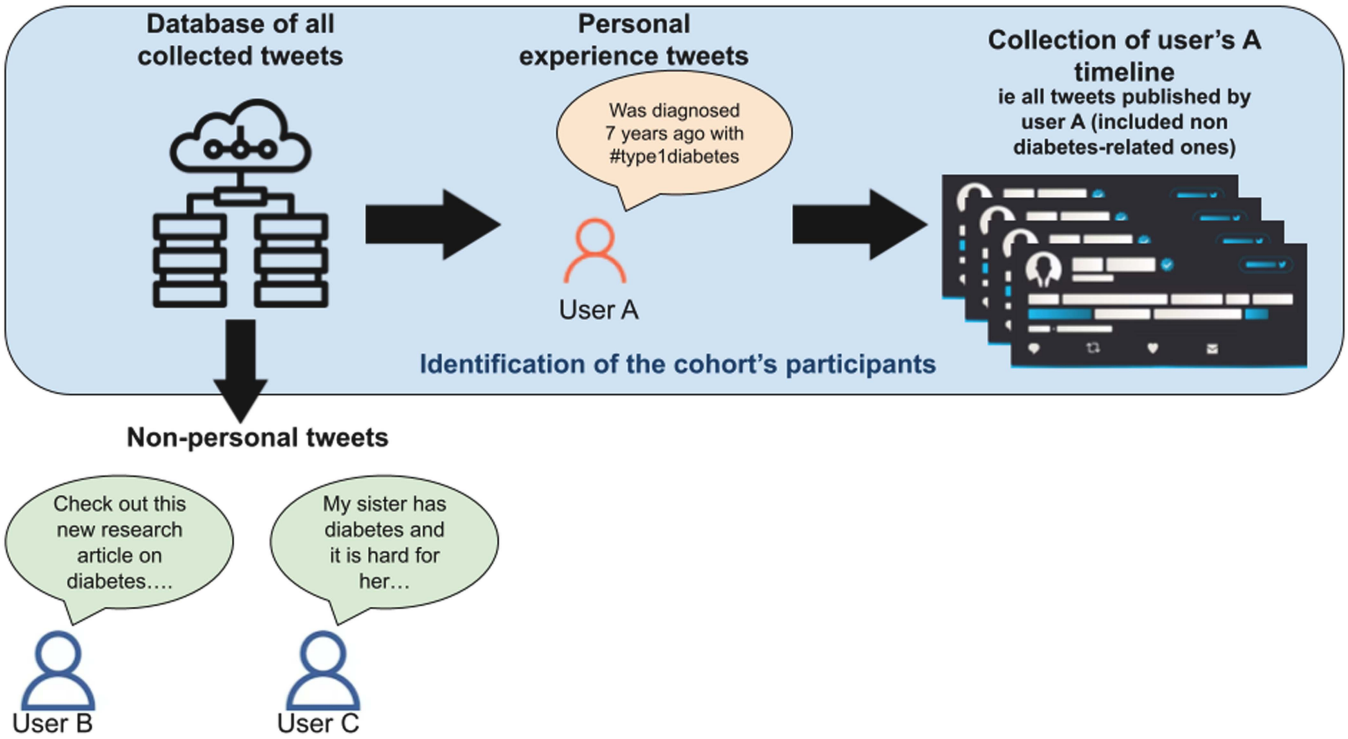

Fig. 2. Identification of users tweeting about their personal experience with diabetes and collection of their timelines.

TABLE II
SIMILARITIES WITH CONCEPT "DIABETES" USING DIFFERENT MODELS

Model	Text	Similarity
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

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 compute 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.

each concept using this model to define manually the thresholds. Thresholds can also be found in Appendix 1, available online. We suggest testing this model first to check if it is suitable for further topics.

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 compute 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.

IV. DISCUSSION

In this article 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. It uses SentenceTransformers pre-trained models to compute semantic similarities with health concepts (exposure and outcomes). For our use case, we relied on Twitter as previous work by Klein et al. [28] 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. It is important to mention that other timelines-based social media data than Twitter can be used to create VDCS with the ALTRUIST package, as long as the data can be formatted and given as input in the workflow as presented in this work. 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 [29]. It provides a way to collect timelines and analyze data using Language Inquiry and Word Count (LIWC) dictionaries which require a paid license [30]. ALTRUIST can be used as a complementary approach to this package once consents, user IDs and potentially collect data to analyze the data.

Our use case, a VDCS on diabetes, apart from the data collection initiated by the World Diabetes Distress Study project [23], 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 quality of life and mental health [31], [32]. Struijs et al. showed that people with diabetes with diabetes-related and non-diabetes-related comorbidities increase the health care demand [33]. We were able to show how people with diabetes-related comorbidities and who tweet about it are more at risk of mentioning 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 [34]. 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 U.K. Biobank study in the United Kingdom includes 500,000 participants [35], and the All of Us Research Program in the USA aims to reach up to 1 million participants [36]. 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 disease. Finally, the best model selection is based on the testing of several existing pre-trained models. The user can 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. Moreover, social media data can pose a challenge regarding the accuracy of health information shared by users. Unlike clinical data, which also has biases such as patients underreporting symptoms, social media data lacks direct verification. To address such issues, we rigorously labeled the data for our "personal content" classifier to minimize false positives and ensure the inclusion of individuals with diabetes. 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, using a GPU is not a mandatory requirement. In our use case, we successfully conducted our analysis on millions of tweets using only a CPU. The choice between CPU and GPU can be made based on the specific requirements of the analysis, particularly the speed of processing desired. In all, 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 the 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.

V. Conclusion

We developed an open-source Python package that facilitates the generation and analysis of VDCS from social media data. It is an easy-to-use tool to add to the arsenal of health researchers to run digital epidemiology projects. This methodology aligns with the principle of minimally disruptive clinical research, prioritizing the participants' well-being. It offers meaningful and comprehensive insights without the need for direct patient involvement. To determine the suitability of the approach for a specific research question, we recommend referring first to the original methodological paper on VDCS. ALTRUIST can be used either as a standalone approach or in conjunction with

traditional research methods, ensuring and providing a comprehensive understanding of patients' experiences throughout the study. In the dynamic field of digital epidemiology, ALTRUIST will keep evolving and broaden its capabilities. Future versions could include increased automated processes.

Acknowledgment

The content of this publication is solely the responsibility of the authors and does not necessarily represent the official views of the funders. GF takes full responsibility for the work as a whole, and for the decision to submit and publish the manuscript. The authors' contributions were as follows: CB and GF designed the research; CB and GF conducted the research; CB collected the data; CB and GF labeled the data; CB and GF analyzed data; CB and GF interpreted the data; CB, GF, and LG developed the package; CB and GF drafted the article; GF, AE, AA, LG revised the manuscript critically.

References

[1] D. Barrett and H. Noble, "What are cohort studies?," *Evid Based Nurs.*, vol. 22, pp. 95–96, 2019.
[2] "Cohort study," Feb. 2013. [Online]. Available: https://www.iwh.on.ca/what-researchers-mean-by/cohort-study
[3] M. S. Setia, "Methodology series module 1: Cohort studies," *Indian J. Dermatol.*, vol. 61, pp. 21–25, 2016.
[4] J. W. Song and K. C. Chung, "Observational studies: Cohort and case-control studies," *Plast Reconstr Surg.*, vol. 126, pp. 2234–2242, 2010.
[5] G. Fagherazzi, C. Bour, and A. Ahne, "Emulating a virtual digital cohort study based on social media data as a complementary approach to traditional epidemiology: When, what for, and how?," *Diabetes Epidemiol. Manage.*, vol. 7, 2022, Art. no. 100085.
[6] C. Bour, A. Ahne, S. Schmitz, C. Perchoux, C. Dessenne, and G. Fagherazzi, "The use of social media for health research purposes: Scoping review," *J. Med. Internet Res.*, vol. 23, 2021, Art. no. e25736.
[7] D. Arigo, S. Pagoto, L. Carter-Harris, S. E. Lillie, and C. Nebeker, "Using social media for health research: Methodological and ethical considerations for recruitment and intervention delivery," *Digit Health*, vol. 4, 2018, Art. no. 2055207618771757.
[8] K. Makice, *Twitter API: Up and Running: Learn How to Build Applications With the Twitter API*, Sebastopol, CA, USA: O'Reilly Media, Inc., 2009.
[9] S. Aslam, "Twitter by the numbers (2023): Stats, demographics & Fun Facts," in *Omnicore Agency*, Mar. 9, 2023. [Online]. Available: https://www.omnicoreagency.com/twitter-statistics/
[10] C. Bour et al., "Global diabetes burden: Analysis of regional differences to improve diabetes care," *BMJ Open Diabetes Res Care*, vol. 10, 2022, Art. no. e003040, doi: 10.2139/ssrn.4128868.
[11] A. Sinha, T. Porter, and A. Wilson, "The use of online health forums by patients with chronic cough: Qualitative study," *J. Med. Internet Res.*, vol. 20, 2018, Art. no. e19.
[12] L. J. Finney Rutten, K. D. Blake, A. J. Greenberg-Worisek, S. V. Allen, R. P. Moser, and B. W. Hesse, "Online Health Information Seeking Among US Adults: Measuring Progress Toward a Healthy People 2020 Objective," *Public Health Rep.*, vol. 134, pp. 617–625, 2019.
[13] A. M. Abu Dabrh, K. Gallacher, K. R. Boehmer, I. G. Hargraves, and F. S. Mair, "Minimally disruptive medicine: The evidence and conceptual progress supporting a new era of healthcare," *J. Roy. College Physicians Edinburgh*, vol. 45, pp. 114–117, 2015.
[14] Drake Van, *Python 3 Reference Manual*, Scotts Valley, CA, USA: CreateSpace, 2009.
[15] N. Reimers and I. Gurevych, "Sentence-BERT: Sentence embeddings using siamese BERT-networks," 2019, *arXiv: 1908.10084*.
[16] S. Bradshaw, E. Brazil, and K. Chodorow, *MongoDB: The Definitive Guide: Powerful and Scalable Data Storage*, Sebastopol, CA, USA: O'Reilly Media, Inc., 2019.
[17] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," 2018, *arXiv: 1810.04805*.
[18] D. Q. Nguyen, T. Vu, and A. T. Nguyen, "BERTweet: A pre-trained language model for English Tweets," 2020, *arXiv: 2005.10200*.
[19] "Deep-translator," Jun. 28, 2023. [Online]. Available: https://pypi.org/project/deep-translator/
[20] "datetime - Basic date and time types," in *Python Documentation*, Dec. 15, 2023. [Online]. Available: https://docs.python.org/3/library/datetime.html
[21] "Survival regression – lifelines 0.27.3 documentation," Oct. 13, 2022. [Online]. Available: https://lifelines.readthedocs.io/en/latest/Survival%20Regression.html
[22] S. L. Spruance, J. E. Reid, M. Grace, and M. Samore, "Hazard ratio in clinical trials," *Antimicrob Agents Chemother*, vol. 48, pp. 2787–2792, 2004.
[23] A. Ahne 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*, vol. 8, 2020, Art. no. e001190, doi: 10.1136/bmjdrc-2020-001190.
[24] D. A. Kiriella et al., "Unraveling the concepts of distress, burnout, and depression in type 1 diabetes: A scoping review," *EClinicalMedicine*, vol. 40, 2021, Art. no. 101118.
[25] K. E. Kreider, "Diabetes distress or major depressive disorder? A practical approach to diagnosing and treating psychological comorbidities of diabetes," *Diabetes Ther.*, vol. 8, pp. 1–7, 2017.
[26] A. N. Long and S. Dagogo-Jack, "Comorbidities of diabetes and hypertension: Mechanisms and approach to target organ protection," *Amer. J. Hypertension*, vol. 13, pp. 244–251, 2011.
[27] "Sentence-transformers/all-mpnet-base-v2. hugging face," 2021. [Online]. Available: https://huggingface.co/sentence-transformers/all-mpnet-base-v2
[28] A. Z. Klein, K. O'Connor, L. D. Levine, and G. Gonzalez-Hernandez, "Using Twitter data for cohort studies of drug safety in pregnancy: Proof-of-concept with β-blockers," *JMIR Formative Res.*, vol. 6, no. 6, 2022, Art. no. e36771.
[29] A. R. Tanner et al., "Epicosm-a framework for linking online social media in epidemiological cohorts," *Int. J. Epidemiol.*, vol. 52, pp. 952–957, 2023, doi: 10.1093/ije/dyad020.
[30] Y. R. Tausczik and J. W. Pennebaker, "The psychological meaning of words: LIWC and computerized text analysis methods," *J. Lang. Social Psychol.*, vol. 29, pp. 24–54, 2010.
[31] M. C. Adriaanse, H. W. Drewes, I. van der Heide, J. N. Struijs, and C. A. Baan, "The impact of comorbid chronic conditions on quality of life in type 2 diabetes patients," *Qual Life Res.*, vol. 25, pp. 175–182, 2016.
[32] S. Pati, S. Pati, M. van den Akker, F. F. G. Schellevis, S. Jena, and J. S. Burgers, "Impact of comorbidity on health-related quality of life among type 2 diabetic patients in primary care," *Primary Health Care Res. Develop.*, vol. 21, 2020, Art. no. e9.
[33] J. N. Struijs, C. A. Baan, F. G. Schellevis, G. P. Westert, and G. A. M. van den Bos, "Comorbidity in patients with diabetes mellitus: Impact on medical health care utilization," *BMC Health Serv Res.*, vol. 6, 2006, Art. no. 84.
[34] S. V. Bădescu et al., "The association between diabetes mellitus and depression," *J. Med. Life*, vol. 9, pp. 120–125, 2016.
[35] J. Elliott et al., "COVID-19 mortality in the UK Biobank cohort: Revisiting and evaluating risk factors," *Eur. J. Epidemiol.*, vol. 36, pp. 299–309, 2021.
[36] All of Us Research Program Investigators et al. "The "All of Us" Research Program," *New England J. Med.*, vol. 381, pp. 668–676, 2019.

Charline Bour is currently working towards the PhD degree with the University of Luxembourg and the Luxembourg Institute of Health (LIH), Luxembourg. Her research interests include the use of artificial intelligence methods to study social media data in digital epidemiology, with a particular focus on natural language processing applied to diabetes-related content.

Abir Elbeji is currently working towards the PhD degree with the University of Luxembourg and the Luxembourg Institute of Health (LIH), Luxembourg. Her research interests include audio signal/speech processing and primarily concentrated on the development of voice-based biomarkers for monitoring patients with severe health conditions.

Adrian Ahne received the PhD degree in AI and digital health from Paris-Saclay University and Inserm. Currently, he is working on the identification of vocal biomarkers for various diseases and their implementation into clinical practice.

Luigi De Giovanni is an IT professional with expertise in data management and analysis. He has experience with a global enterprise for entertainment data and he is now working as IT support for data on research projects with the Luxembourg Institute of Health (LIH).

Guy Fagherazzi is a director with the Department of Precision Health, Luxembourg Institute of Health (LIH). He is also the head of the Deep Digital Phenotyping Research Unit, a multidisciplinary research lab with LIH where they conduct data-driven digital phenotyping research using large cohort studies and online data to improve the understanding of the impact of various chronic diseases (diabetes, mental health, cancer, Long COVID) on the daily lives of patients and populations.