International Journal of Epidemiology, 2023, 1–6
https://doi.org/10.1093/ije/dyad020
Software Application Profile

Software Application Profile

Epicosm—a framework for linking online social media in epidemiological cohorts

Alastair R Tanner [1], Nina H Di Cara,[1,2] Valerio Maggio,[1,2] Richard Thomas,[2] Andy Boyd,[2] Luke Sloan,[4] Tarek Al Baghal,[5] John Macleod,[2] Claire M A Haworth,[3,6†] and Oliver S P Davis[1,2,6,*]

[1]Medical Research Council Integrative Epidemiology Unit, University of Bristol, Bristol, UK, [2]Department of Population Health Sciences, Bristol Medical School, University of Bristol, Bristol, UK, [3]School of Psychological Science, University of Bristol, Bristol, UK, [4]School of Social Sciences, Cardiff University, Cardiff, UK, [5]Institute for Social and Economic Research, University of Essex, Colchester, UK and [6]Alan Turing Institute, British Library, London, UK

*Corresponding author. Department of Population Health Sciences, Bristol Medical School, University of Bristol, Bristol BS8 2BN, UK. E-mail: oliver.davis@bristol.ac.uk

[†]Co-supervisors.

Received 7 February 2022; Editorial decision 16 January 2023; Accepted 16 February 2023

Abstract

Motivation: Social media represent an unrivalled opportunity for epidemiological cohorts to collect large amounts of high-resolution time course data on mental health. Equally, the high-quality data held by epidemiological cohorts could greatly benefit social media research as a source of ground truth for validating digital phenotyping algorithms. However, there is currently a lack of software for doing this in a secure and acceptable manner. We worked with cohort leaders and participants to co-design an open-source, robust and expandable software framework for gathering social media data in epidemiological cohorts.

Implementation: Epicosm is implemented as a Python framework that is straightforward to deploy and run inside a cohort's data safe haven.

General features: The software regularly gathers Tweets from a list of accounts and stores them in a database for linking to existing cohort data.

Availability: This open-source software is freely available at [https://dynamicgenetics.github.io/Epicosm/].

Key words: Social media, epidemiology, cohort studies, longitudinal studies, data science, Big Data, mental health, wellbeing, data linkage, ALSPAC

Introduction

Digital footprint data, such as data from social media, banking and shopping, online searches, and apps such as exercise trackers, offer huge potential for epidemiological studies to derive new digital phenotypes based on real human behaviour. For example, research predicting mental health from digital data has been increasing since 2013,[1] and progress in this area could improve access to mental health care, such as by offering overstretched services a way of supporting patients between check-in occasions. Early detection of problems through digital phenotyping could lead to early interventions that prevent the development of more complex issues. This might work at both an individual level[2] and a strategic level, where services can be put in place to meet anticipated demands of populations such as student groups,[3,4] emergency workers[5] or geographical regions,[6] an approach that proved particularly valuable during the COVID-19 pandemic.[7–10] Research in this field has developed methods for inferring a wide range of outcomes, including social anxiety,[11] suicidality,[12,13] depression,[14–16] wellbeing[6,17] and happiness.[18,19] This makes social media data a potentially valuable source of information for epidemiological studies, such as birth cohorts, to supplement more traditional approaches. Inference from social media data has the potential to provide high temporal resolution data on mental wellbeing on daily or even hourly timescales, and research using these data could advance our understanding of mental health time courses, aid early diagnosis and inform public health interventions and policies,[20,21] opportunities recognized by funding bodies, such as United Kingdom Research and Innovation (UKRI) and the Wellcome Trust, in their cohort data linkage strategies.[22,23]

Conversely, researchers developing approaches for inferring phenotypes from these novel data can benefit from the resources provided by epidemiological cohorts. Historically, studies using social media have rarely had good knowledge of the samples they were studying, risking demographic bias and unmeasured confounding. Similarly, these studies have rarely had access to good 'ground truth' measures of their phenotypes of interest. Epidemiological cohorts, with their well characterized participants and state-of-the-art phenotyping, offer an opportunity for a step change in the quality of research in this area by allowing straightforward validation of new digital measures against gold-standard, symptom-based assessments and diagnoses in a known population.

Despite these advantages, currently few longitudinal cohorts have linked digital footprint data, because of the specific challenges. For example, social media data are difficult to anonymize; with a publicly available platform such as Twitter, there is no way for a cohort to share user names or Tweets without identifying cohort participants. This is particularly important because cohorts rely on a long-term trust relationship with participants, and the disclosure of personal data could lead to reputational damage for the cohort and a decrease in participation. Such challenges mean that cohorts could benefit greatly from software designed in collaboration with cohort leaders and participants to meet their specific needs. With this in mind, we worked closely with stakeholders from the Avon Longitudinal Study of Parents and Children (ALSPAC) and other CLOSER (based at the University College London Social Research Institute) cohorts to design the Epicosm software to address these special requirements.

Although social media harvesting software products are widely available, most require significant programming skills run in a way useful to longitudinal cohorts, such as collecting new data regularly from a list of specific users. Similarly, most social media harvesters are not well documented, and do not provide functions such as data management or built-in approaches for inferring common digital phenotypes from datasets. In contrast, Epicosm is designed to be relatively straightforward to set up and run on servers in the heterogeneous computing environments inside cohorts' data safe havens, allowing long-term linking of social media time lines from a list of users, storage of information in a flexible database structure, and automated and modular processing of the data using several widely used coding algorithms. At the time of writing, Epicosm's focus is on harvesting data from Twitter. However, the software has been designed following software engineering best practices, including modular organization to allow expansion to other social media platforms, Open Source code available on GitHub, and documentation written with future collaborators and maintainers in mind. The data collected by Epicosm form the basis for a depersonalized dataset of information, derived from social media, which can be shared with researchers through a cohort's usual data access mechanisms. As social media big data continue to evolve, Epicosm provides robust data acquisition tools so that epidemiology can benefit from these rich sources of information about the daily lives and behaviours of people and populations.

Implementation

Epicosm is an open-source project freely available under a GPL version 3 licence from GitHub [dynamicgenetics.github.io/Epicosm/], along with full documentation. Collaborators are welcome to branch, fork or issue pull requests [for example, updating in response to changing API (Application Programming Interface) authentication]

and to add custom functionality. The modular nature of the software suite allows adaptation for alternative platforms, allowing any typical API response to be archived in a local database for later analysis. Together, the software engineering principles applied in its development promote collaboration to maintain and expand Epicosm's scope, and allow it to act as a foundation for a variety of research.

For data management, Epicosm uses the open-source, non-relational document database MongoDB [mongodb.com]. MongoDB was chosen for flexibility: the schema is consistent with Java Script Object Notation (JSON) data structure, a common format for API responses. This allows the storage of a variety of types of data (from plain text with metadata to images and other media), and accommodates adaptation of Epicosm to variation in API responses over time and across a range of social media platforms.

We anticipate that Epicosm will be installed and managed by cohort data managers: these staff typically have permissions to process identifiable participant data and are responsible for the post-processing (for example de-personalization) needed prior to sharing with researchers. In development we have been sensitive to user requirements, keeping requisite skills marginal: some basic experience of the command line interface is expected, but no programming experience is required and we provide full instructions for setting up and running the software. The repository also contains links to resources to support new users.

The steps to gather information from the Twitter API using Epicosm are as follows. The user must provide two files: (i) a list of participants' Twitter user names (also known as 'screen names' or 'handles'); and (ii) a Twitter API bearer token to authorize API requests. Once these are in place, Epicosm is ready to run and carries out the following processes.

i. Credentials are verified by Twitter's API.
ii. The API converts screen names to unique and persistent identification (ID) numbers (this enables the tracking of participant accounts longitudinally, even where participants change screen names).
iii. Epicosm then requests Twitter timelines—that is, the user's tweets (posts by the user) and re-tweets (re-posts of other users' posts)—from each ID number. With an authorised academic research account, the complete tweet history of each user is available.
iv. Finally, each record (a single JSON document for each tweet) is stored in MongoDB. The tweet harvest can be scheduled to repeat at regular intervals specified by the user.

Various options are available, depending on the specific consent obtained from participants, including acquiring the list of 'followed', third-party Twitter account names. Public followed accounts can also be harvested for their tweets: the content of this harvest approximates the 'feed' that a user is presented with by Twitter (or at least, the pool of tweets available for Twitter to present to the user). In contrast to the original user tweet harvest, this harvest will only acquire posts made in the last 7 days (but can be repeated weekly). A full-archive harvest of followed accounts is theoretically possible, but not currently implemented in Epicosm: users can each follow thousands of accounts, each of which may have a large history of tweets, especially if they are intensely managed (for example, celebrity accounts) or automated accounts (for example, sports results or the weather).

Epicosm includes a selection of widely used algorithms (Box 1) for deriving sentiment (and other) information from Twitter data: LabMT,[18] VADER,[24] LIWC2015[25] and TextBlob[26] (note that, for licensing reasons, LIWC analysis requires the users to acquire a dictionary from the LIWC developers). Epicosm applies the analyses to each tweet, and appends these to each record in labelled database fields. The software provides implementations of these commonly used measures, as a demonstration of how phenotypes can be automatically added to the data base and because they are likely to be requested by researchers. However, the platform is flexible to allow users to derive novel phenotypes through the addition of custom algorithms or new dictionaries to allow analysis of languages other than English, and we anticipate that cohorts will employ a variety of their own approaches to derive information from the Twitter data once Epicosm has downloaded and stored it.

Use

As an approximate guide based on a random sample, 1000 users typically have an acquirable history of around 700 000 tweets, leading to a database size of around 3.5 GB, although the software will also allow data collection from much larger samples, limited only by storage space and the Twitter API's rate limits. When first run, Epicosm will attempt to gather the full tweet time line history for each user. Subsequent harvesting operations will return only the tweets more recent than the latest tweet already in the database. At the time of writing, data can be acquired at about a million tweets per hour, but this will be highly dependent on connection speed, network activity and any rate-limiting measures Twitter impose (i.e. where they restrict the rate of download via the API).

As an example of expected use, we gathered tweets from a list of around 800 Twitter accounts of consenting participants in ALSPAC, an epidemiological birth cohort of around 15 000 families recruited between 1991 and 1992 in the historical county of Avon in the west of the UK.[27] ALSPAC was interested in understanding the potential of these novel data to infer changes in mental health over time. Participants provided informed consent and ethical approval was provided by the ALSPAC Ethics and Law committee. We used Epicosm to link Twitter data as proof-of-principle. Of course, different populations use social media platforms in different ways, and this evolves with time. Twitter users, in the UK at least, are on average younger and slightly more likely to be male than the general population,[28] although there is less age bias than previously assumed, with good representation from all age groups. For the ALSPAC young people at 24 years old, there was little difference in Twitter use across gender, ethnicity and parental employment groups, although those who had completed Advanced Level qualifications (post-16 school leaving examinations) were slightly more likely to use Twitter (58% compared with 51%).[29] Despite the potential for bias in the sample, cohorts are the ideal for collecting this type of data because the biases are often identifiable. Twitter is currently among the social media platforms most open to academic research, but our intention is to expand Epicosm's capabilities in future to include linking other forms of social media, subject to API restrictions.

The Twitter data linkage in ALSPAC was guided by conversations with cohort participants to understand the acceptability of this use of data and to establish appropriate safeguards,[30] and with cohort data managers and linkage experts to understand the requirements for running the software and retrieving the data. These insights emphasize the wider evidence from participants[30,31] that it is a necessity for data accessed by researchers to be depersonalized, and that study data managers operate in a trusted role where they are able to capture identifiable data and process these so they are suitable for dissemination to researchers.

We developed a data management protocol that ensured that Twitter linkage fitted with ALSPAC's linkage data pipeline model, acquiring tweets from consenting participants via the Twitter Application Programming Interface (API) and depositing these in raw form in a permanent, versioned MongoDB data base. Data bases such as MongoDB are particularly useful for social media data because they store data in the form of documents that are very similar to the responses received from social media APIs. In this case, each document corresponded to a tweet and its associated metadata. The ALSPAC data managers (who have exclusive access to participant identifiers) followed a protocol that involved:

i. the implementation of consent and withdrawal;
ii. providing the software with a list of Twitter user names which guided the collection of data from the Twitter API;

Box 1 Sentiment analysis methods

Sentiment analysis, the most common approach applied to derive information from social media data, is the inference of emotions, opinions and attitudes from written text. Output metrics vary depending on the methodology, but common inferences include positive or negative emotions or a composite of both [for example, VAD ER[29] (Valence Aware Dictionary and sEntiment Reasoner)]. Some methods also aim to derive more specific emotional and syntactic content, for example LIWC[30] (Linguistic Inquiry and Word Count) infers over 70 categories from emotions aor gender specificities to politics or food.

A commonly used methodology is the 'dictionary approach'. Individual words are first assigned sentiment scores by a group of participants, to build up a dictionary. For example, words such as 'death', 'hate' and 'hell' might be assigned negative scores, and 'friend', 'happy' and 'love' are generally rated more positively. The text is then assigned a mean score based on the dictionary words it contains (or a relative frequency for categorical dictionaries). This straightforward approach is limited by features of natural language such as negation, neologism, irony or sarcasm, but these are often equally difficult for human readers to understand, and their influence can be mitigated by applying more sophisticated natural language processing and machine learning approaches that aim to interpret sentence structure or do not assume the direct correspondence between the dictionary definition of words and the associated phenotypes. Linking social media data in epidemiological cohorts provides a crucial tool to develop these new approaches, by providing access to linked independent outcome ('ground truth') measures and demographic information about the populations studied.

iii. subsequent curation of the data in the MongoDB database (including documentation and versioning);
iv. the management of participant identifiers to enable linkage to other cohort data;
v. ensuring that the raw captured data were sufficiently depersonalized to share as structured data outputs, while retaining full Twitter content within the cohort's data safe haven for the duration of the study so that it could be repurposed for future research needs.

Conclusion

Social media data offer huge potential for digital phenotyping in epidemiological cohort studies to complement traditional measures. Equally, epidemiological cohorts have much to offer digital footprint researchers. We have described the software Epicosm and how it can be used by epidemiologists to expand existing cohort datasets. The software provides a robust foundation for Twitter data acquisition, and enables the exploration of participants' digital footprints to address important health and social research questions. As the importance of online community and communication increases (especially in light of global health events such as the COVID-19 pandemic), Epicosm offers epidemiologists a practical way to expand their work into novel types of data and methodologies, and opens up the valuable data already held by longitudinal population cohorts to new research communities.

Ethics approval

This study and all related work were approved by the ALSPAC Ethics and Law Committee (Haworth.Davis.B2934).

Data availability

This Open Source software is freely available at [https://dynamicgenetics.github.io/Epicosm/].

Author contributions

O.S.P.D. and C.M.A.H. conceived the program structure, acquired the funding with L.S., T.A.B., A.B. and J.M., authored the manuscript and supervised the project. A.R.T. designed and developed Epicosm code and authored the manuscript. N.D.C. and V.M. contributed to coding, testing and authoring. All other authors contributed to developing the wider framework for using social media data within longitudinal studies and to writing the manuscript. The data management and design of the data pipelines within the ALSPAC cohort were conducted by R.T. and managed by A.B. and J.M.

Funding

This project is funded by CLOSER [www.closer.ac.uk], whose mission is to maximize the use, value and impact of longitudinal studies. CLOSER was funded by the Economic and Social Research Council (ESRC) and the Medical Research Council (MRC) between 2012 and 2017. Its initial 5-year grant has since been extended to March 2021 by the ESRC (grant reference: ES/K000357/1). The funders took no role in the design, execution, analysis or interpretation of the data or in the writing up of the findings. The UK Medical Research Council and Wellcome Trust (Grant ref: 217065/Z/19/Z) and the University of Bristol provide core support for ALSPAC. This publication is the work of the authors, and A.R.T. and O.S.P.D. will serve as guarantors for the contents of this paper. A comprehensive list of ALSPAC grant funding is available on the ALSPAC website [http://www.bristol.ac.uk/alspac/external/documents/grant-acknowledgements.pdf]. The work was supported by the UK Medical Research Council through the ALSPAC and Born In Bradford Mental Health Pathfinder award (grant reference: MC_PC_17210), and in part by the UK Medical Research Council Integrative Epidemiology Unit at the University of Bristol (Grant ref: MC_UU_12013/1). O.S.P.D. and C.M.A.H. .are funded by the Alan Turing Institute under the EPSRC grant EP/N510129/1. C.M.A.H. is supported by a Philip Leverhulme Prize.

Acknowledgements

A.R.T. would like to thank Chris Edsall, Christopher Woods and the research software engineering team at the University of Bristol. We are extremely grateful to all the ALSPAC families who took part in the proof-of-principle study, the midwives for their help in originally recruiting them, and the whole ALSPAC team, which includes interviewers, computer and laboratory technicians, clerical workers, research scientists, volunteers, managers, receptionists and nurses.

Conflict of interest

None declared.

References

1. De Choudhury M, Gamon M, Counts S *et al.* Predicting depression via social media. In: *Seventh International AAAI Conference on Weblogs and Social Media, 8–11 July 2013*. Cambridge MA: Association for the Advancement of Artificial Intelligence, 2013.
2. McGorry PD, Ratheesh A, O'Donoghue B. Early intervention—an implementation challenge for 21st century mental health care. *JAMA Psych* 2018;**75**:545–46.
3. Melcher J, Lavoie J, Hays R *et al.* Digital phenotyping of student mental health during COVID-19: An observational study of 100 college students. *J Am Coll Health* 2021;**Mar 26**:1–13.
4. Melcher J, Hays R, Torous J. Digital phenotyping for mental health of college students: a clinical review. *Evid Based Ment Health* 2020;**23**:161–66.
5. Blair J, Hsu C-Y, Qiu L *et al.* Using tweets to assess mental well-being of essential workers during the covid-19 pandemic. In: *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems,, 8–13 May 2021*. Yokohama, Japan. New York, NY: Association for Computing Machinery, 2021, pp. 1–6.

6. Jaidka K, Giorgi S, Schwartz HA *et al.* Estimating geographic subjective well-being from twitter: a comparison of dictionary and data-driven language methods. *Proc Natl Acad Sci U S A* 2020;**117**:10165–71.
7. Di Cara NH, Song J, Maggio V *et al.* Mapping population vulnerability and community support during COVID-19: a case study from Wales. *Int J Popul Data Sci* 2021;**5**:1409.
8. Pellert M, Lasser J, Metzler H *et al.* Dashboard of sentiment in Austrian social media during COVID-19. *Front Big Data* 2020; **3**:32.
9. Guntuku SC, Buffone A, Jaidka K, *et al.* Understanding and measuring psychological stress using social media. In: *Proceedings of the International AAAI Conference on Web and Social Media, 11–14 June 2019*, Munich, Germany. Cambridge, MA: Association for the Advancement of Artificial Intelligence, 2019, pp. 214–25.
10. Xue J, Chen J, Hu R *et al.* Twitter discussions and emotions about the COVID-19 pandemic: Machine learning approach. *J Med Internet Res* 2020;**22**:e20550.
11. Lee J, Sohn D, Choi YS. A tool for spatio-temporal analysis of social anxiety with twitter data. In: *Proceedings of the 34th ACM/SIGAPP Symposium on Applied Computing, 8–12 April, 2019*, Limassol, Cyprus. New York, NY: Association for Computing Machinery, pp. 2120–23.
12. Choi D, Sumner SA, Holland KM *et al.* Development of a machine learning model using multiple, heterogeneous data sources to estimate weekly US suicide fatalities. *JAMA Netw Open* 2020;**3**:e2030932.
13. Sinyor M, Williams M, Zaheer R *et al.* The association between twitter content and suicide. *Aust N Z J Psychiatry* 2021;**55**: 268–76.
14. Cohrdes C, Yenikent S, Wu J *et al.* Indications of depressive symptoms during the COVID-19 pandemic in Germany: comparison of national survey and twitter data. *JMIR Ment Health* 2021;**8**:e27140.
15. Li D, Chaudhary H, Zhang Z. Modeling spatiotemporal pattern of depressive symptoms caused by COVID-19 using social media data mining. *IJERPH* 2020;**17**:4988.
16. Sharma S, Sharma S. Analyzing the depression and suicidal tendencies of people affected by COVID-19's lockdown using sentiment analysis on social networking websites. *J Stat Manage Syst* 2021;**24**:115–33.
17. Zhang X, Wang Y, Lyu H *et al.* The influence of COVID-19 on the well-being of people: Big data methods for capturing the well-being of working adults and protective factors nationwide. *Front Psych* 2021;**12**:2327.
18. Dodds PS, Harris KD, Kloumann IM *et al.* Temporal patterns of happiness and information in a global social network: hedonometrics and twitter. *PLoS One* 2011;**6**:e26752.
19. Al Shehhi A, Thomas J, Welsch R *et al.* Arabia felix 2.0: a cross-linguistic twitter analysis of happiness patterns in the United Arab Emirates. *J Big Data* 2019;**6**:1–20.
20. Torous J, Baker JT. Why psychiatry needs data science and data science needs psychiatry: connecting with technology. *JAMA Psychiatry* 2016;**73**:3–4.
21. Guntuku SC, Yaden DB, Kern ML, Ungar LH, Eichstaedt JC. Detecting depression and mental illness on social media: an integrative review. *Curr Opin Behav Sci* 2017;**18**:43–49.
22. UK Research and Innovation. *MRC Strategic Review of the Largest UK Population Cohort Studies*. https://mrc.ukri.org/publications/browse/maximising-the-value-of-uk-population-cohorts/ (December 2022, date last accessed).
23. The Wellcome Trust. *Enabling Data Linkage to Maximise the Value of Public Health Research Data: Full Report*. https://wellcome.org/sites/default/files/enabling-data-linkage-to-maximise-value-of-public-health-research-data-phrdf-mar15.pdf (December 2022, date last accessed).
24. Gilbert CHE, Hutto E. Vader: A parsimonious rule-based model for sentiment analysis of social media text. http://comp social gatech edu/papers/icwsm14 (December 2022, date last accessed).
25. Pennebaker JW, Boyd RL, Jordan K, Blackburn K. *The Development and Psychometric Properties of LIWC2015*. 2015. https://repositories.lib.utexas.edu/handle/2152/31333 (December 2022, date last accessed).
26. Textblob. pypi.org/project/textblob/ (December 2022, date last accessed).
27. Boyd A, Thomas R, Hansell AL *et al.* Data resource profile: the ALSPAC birth cohort as a platform to study the relationship of environment and health and social factors. *Int J Epidemiol* 2019; **48**:1038–39k.
28. Sloan L. Who tweets in the United Kingdom? Profiling the Twitter population using the British social attitudes survey 2015. *Social Media Soc* 2017;**3**. doi: 10.1177/2056305117698981.
29. Di Cara NH, Winstone L, Sloan L, Davis OSP, Haworth CMA. The mental health and well-being profile of young adults using social media. *NPJ Mental Health Res* 2022;**1**:11.
30. Di Cara NH, Boyd A, Tanner AR *et al.* Views on social media and its linkage to longitudinal data from two generations of a UK cohort study. *Wellcome Open Res* 2020;**5**:44.
31. Sloan L, Jessop C, Al Baghal T, Williams M. Linking survey and Twitter data: informed consent, disclosure, security, and archiving. *J Empir Res Hum Res Ethics* 2020;**15**:63–76.