### DATA SHARING AGREEMENT

### (Self-reported bipolar disorder diagnosis (S-BiDD) dataset)

### between

### UNIVERSITY OF LANCASTER

### and

### INSERT NAME OF RECIPIENT INSTITUTION

# DATA SHARING AGREEMENT (**S-BiDD dataset**)

between

Lancaster University, an educational body incorporated by Royal charter and having its main administrative offices at University House, Bailrigg, Lancaster, LA1 4YW (the “Provider Institution”)

and

Insert full name of recipient institution, a charitable body registered in Insert jurisdiction of registration under registration number Insert charitable or royal charter number, incorporated under Insert Act or Charter of incorporation and having its main administrative offices at Insert full legal address of recipient institution (“Recipient Institution”)

hereinafter referred to as “the Parties” and each of them being “a Party”

# BACKGROUND

1. The Recipient Institution is conducting a research project entitled “Insert title of research project” as described in more detail at Schedule 1 (the “Approved Research”) under the direction of Insert PI (“the Recipient Scientist”) and wishes to access and use the data specified in Schedule 2 (the “Data”) for the purpose of the Approved Research.
2. The Provider Institution is willing to supply the Data to the Recipient Institution and the Recipient is willing to receive the Data in accordance with the terms and conditions contained within this agreement (the “Agreement”).

# TERMS AND CONDITIONS

It is hereby agreed as follows:

## In this Agreement:

## the term “Data” includes any Data that has been combined or aggregated (wholly or in part) with any other data or information by the Recipient Institution;

1. the term “Approved Research” includes the publication (if any) of the results of the Approved Research by the Recipient Institution;
2. the terms “personal data”, “controller”, “processes”, “processing” and “processed” shall have the meanings given in the Data Protection Laws;
3. “Data Protection Laws” means all applicable data protection and privacy legislation in force from time to time in the UK, including but not limited to: (i) the General Data Protection Regulation ((EU) 2016/679) as retained in UK law by the EU (Withdrawal) Act 2018 and the Data Protection, Privacy and Electronic Communications (Amendments etc) (EU Exit) Regulations 2019 (the "UK General Data Protection Regulation" or “UK GDPR”); (ii) the Data Protection Act 2018; (iii) the Privacy and Electronic Communications Directive 2002/58/EC (as updated by Directive 2009/136/EC) and the Privacy and Electronic Communications Regulations 2003 (SI 2003/2426) as amended; and (iv) all applicable legislation and regulatory requirements in force from time to time which apply to a Party relating to the processing of personal data and privacy and the guidance and codes of practice issued by the Information Commissioner’s Office which apply to a Party.

## In consideration of the obligations accepted by the Recipient Institution under this Agreement, the Provider Institution grants to the Recipient Institution for the duration of the Approved Research a non-exclusive, personal and non-transferable licence to use the Data for the Approved Research.

## The Recipient Institution undertakes to the Provider Institution:

## to use the Data solely for the Approved Research;

1. to restrict access to the Data to the Recipient Scientist and those staff and students comprising the Recipient Scientist’s research team named in the Approved Research proposal in Schedule 1, and to ensure that those staff and students are aware of and comply with the terms of this Agreement;

## to keep the Data confidential and not sub-license, transfer, disclose or otherwise make available the Data in whole or part to any third party except with specific prior written consent from the Provider Institution;

1. to keep the Data secure by implementing organisational and technological measures appropriate to the nature and sensitivity of the data to prevent the unauthorised or accidental access, use or disclosure of the Data, including (without limitation) the measures set out in clause 4;
2. to notify the Provider Institution as soon as reasonably practicable after becoming aware of any unauthorised or accidental access, use or disclosure of the Data, and to co-operate with any investigation made by the Provider Institution in connection with the unauthorised or accidental access, use or disclosure of the Data;

## not attempt to re-identify any individual from the Data or communicate with any individual re-identified from the Data, nor to link or attempt to link the Data to other data or information except with specific prior written consent from the Provider Institution;

## to process the Data in accordance with all applicable laws and regulations [and any regulator’s code(s) of practice applicable to the Data];

1. to permanently delete all copies of the Data from its hard drives and movable media and destroy all physical copies of the Data as soon as reasonably practicable on completion of the Approved Research or on termination of this Agreement (if earlier).

## In particular (and without limiting the generality of the preceding wording) the Recipient Institution shall not make physical or electronic copies of the Data except to the extent reasonably necessary for the Approved Research and shall ensure that any copy of the Data stored on movable media (including laptops) is password-protected and fully encrypted and that any copy of the Data stored on networked or non-networked hard drives is properly protected with firewall and controlled access permissions. The Recipient Institution shall keep a record of where each copy of the Data is stored and shall provide the Provider Institution with a copy of this record on request.

## The Data is supplied by the Provider Institution in pseudonymised form without the pseudonymisation key or other means for the Recipient Institution to re-identify individuals from the Data. The Parties anticipate that the Data is likely to be personal data in respect of the Provider Institution’s processing but not likely to be personal data in respect of the Recipient Institution’s processing, but that this is a question of fact determined by the nature of the Data, the arrangements between the Parties, and any other means available to the Recipient Institution (whether publicly available or otherwise) to re-identify individuals form the Data. In the event that the Data is or becomes personal data when held or processed by the Recipient Institution, the Parties agree that:

## each of them shall be a controller in respect of its own processing of the Data, and shall be solely responsible and liable for its own processing of the Data including (without limitation) the lawful basis for that processing and ensuring that the Data is processed in compliance with the Data Protection Laws; however

1. if the Parties are deemed to be joint controllers in relation to the Data, the Parties shall be jointly responsible and liable for the processing of the Data in connection with this Agreement and shall co-operate to determine their respective responsibilities for compliance with their obligations and duties under the Data Protection Laws.

## 

## Except to the extent prohibited by law, the Recipient Institution assumes all direct liability for damages which may arise from its receipt, use, storage or destruction of the Data. The Provider Institution will not be liable to the Recipient Institution for any use made of the Data, including any loss, claim or demand made by the Recipient Institution or made against the Recipient Institution by a third party, due to or arising from the use, storage or disposal of the Data by the Recipient Institution, except to the extent permitted by law when caused by the gross negligence or wilful misconduct of the Provider Institution.

## The Provider Institution makes no representation and gives no warranty of any kind, either express or implied, including but not limited to warranties of accuracy or fitness for a particular purpose, or that the use of the Data will not infringe any patent, copyright, trademark or other proprietary rights. Data is often contributed by third parties and cannot be guaranteed to be free from errors, omissions or inaccuracies, accordingly the Provider Institution will not be liable to the Recipient Institution for any loss arising from any reliance placed on the Data by the Recipient Institution.

## Nothing in this Agreement limits or excludes either party’s liability for (a) death or personal injury resulting from negligence; or (b) any fraud or for any sort of other liability which, by law, cannot be limited or excluded.

## The liability of either Party for any breach of this Agreement, or arising in any other way out of the subject matter of this Agreement, will not extend to loss of business, or profit, or to any indirect or consequential damages or losses.

## The Data is provided and the Approved Research is undertaken in pursuit of the primary charitable objectives of the Parties; that is the advancement of education through research and teaching. The Provider Institution acknowledges that the results of the Approved Research shall belong to the Recipient Institution (except to the extent the results incorporate or include Data), and that the Recipient Institution may seek to publish the results of the Approved Research. The Recipient Institution shall procure that in relation to any publication reporting on the use of the Data, the Recipient Scientist acknowledges the Provider Institution as the source of the Data in the publication [and acknowledges the funder that funded the collection and compilation of the Data, in the form specified by the Provider Institution or the funder itself]. The Recipient Institution shall provide a copy of such publication to the Provider Institution thirty (30) days in advance of submission for publication, to allow the Provider Institution (as sponsor of the original study under which the Data were collected) an opportunity to comment on the publication. The Provider Institution agrees not to disclose any results contained in such advance copy to any third party until published by the Recipient Institution. The Recipient Institution shall not publish any confidential or proprietary information belonging to the Provider Institution without its prior written consent, including the Data in whole or in part. Confidential and proprietary information shall be deemed to include information which was described as such at the point of disclosure and/or was marked as either “confidential” or “proprietary”. The confidentiality obligations in this clause shall not apply where the confidential or proprietary information:

1. has become public knowledge, other than through an unauthorised disclosure by the Recipient Institution;
2. was already known to the Recipient Institution, prior to disclosure by the Provider Institution;
3. was disclosed to the Recipient Institution or the Recipient Scientist by a third party, whom to the Recipient Institution’s knowledge, was not under any obligation of confidence to the Provider Institution;
4. was released from confidential status by written authorisation of the Provider Institution; or
5. is required to be disclosed by law or by requirement of a regulatory body or court order.

## Recipients shall not publish any verbatim quotes from the dataset but are allowed to publish paraphrased quotes that do not lead to the original online post via a verbatim online search.

## Nothing in this Agreement grants the Recipient Institution any rights over the Data or under any patents, nor any right to use, or permit the use of, any products or processes containing the Data for any profit-making or commercial purposes (“Commercial Use”). Should the Recipient Institution wish to make Commercial Use of the Data and should the Provider Institution be willing and able to grant a licence for such purposes, the Parties shall negotiate in good faith to agree an appropriate licence or revenue sharing agreement on fair and reasonable terms.

## Nothing in this Agreement shall prevent or impede the Provider Institution from being able to use the Data for any purpose, including but not limited to sharing and licensing of the Data to third parties, whether public, private or third sector, for any purpose.

## The rights and obligations of the Parties are personal and may not be assigned at any time without the prior written consent of the other Party which consent shall not be unreasonably withheld; provided that it shall be a requirement in all cases of assignation that the assignee undertakes to perform all outstanding obligations of the assignor as though the assignee had been an original party hereto.

## This Agreement shall be effective from the date of signature and shall continue in force until [TBC] The term of this Agreement may be extended by the mutual written agreement of both Parties signed by their authorised signatories.

## The Provider Institution may terminate this Agreement if the Recipient Institution is in breach of any of the terms of this Agreement and, where the breach is capable of remedy, the Recipient Institution has failed to remedy the same within twenty-eight (28) calendar days of service of a written notice from the Provider Institution specifying the breach and requiring it to be remedied.

## The Data is provided at no cost.

## The Parties shall procure that in carrying out their obligations under this Agreement, they will comply with all applicable laws, regulations and statutes, including those relating to modern slavery and anti-bribery and the Data Protection Laws. Non-compliance with this clause by a Party shall not be sufficient justification for another Party not to comply with its obligations under this Agreement.

## A person who is not a party to this Agreement shall not have any rights under or in connection with it.

## Notices

## The Provider Institution’s representative for the purpose of receiving notices shall until further notice be:

## Director of Research, Enterprise and Innovation

with an email copy to Professor Steven Jones s.jones7@lancaster.ac.uk

## The Recipient Institution’s representative for the purpose of receiving notices shall until further notice be:

## Insert details

## with a copy to:

## Insert details

## This Agreement constitutes the entire agreement between the parties in respect of its subject matter and no statements or representations made by any Party have been relied upon by the other in entering into this Agreement.

## This Agreement shall be governed and construed in accordance with the laws of England and Wales and the Parties agree to the exclusive jurisdiction of the English Courts.

## This Agreement may be executed in one (1) or more counterparts, each of which shall be deemed an original, but all of which together shall constitute one and the same instrument. A signed copy of this Agreement delivered by e-mailed portable document format file or other means of electronic transmission shall be deemed to have the same legal effect as delivery of an original signed copy of this Agreement.

IN WITNESS WHEREOF this Agreement is executed as follows:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| for and on behalf of Lancaster University | |  | for and on behalf of insert name of Recipient Institution | |
| Signed: |  |  | Signed: |  |
| Name: |  |  | Name: |  |
| Title: |  |  | Title: |  |
| Dated: |  |  | Dated: |  |

Schedule 1

**The Approved Research**

[INSERT RESEARCH PLAN/DESCRIPTION Add justification for variables if requesting “User characterisations” version of the dataset release in addition to “Standard”]

Schedule 2

**The Data**

The Self-reported bipolar disorder diagnosis (S-BiDD) dataset contains the public Reddit posts and user meta data of 19,685 Reddit user accounts who self-reported a bipolar disorder on the online discussion platform between 01/2005-03/2019. The dataset is distributed in the parts described below. Please tick all parts that you request access to.

|  |  |
| --- | --- |
| Dataset | Please cite these references if you use this dataset |
| * posts\_meta.csv | (Jagfeld, Lobban, Davies, *et al.*, 2021, under review) |
| * posts\_texts.csv | (Jagfeld, Lobban, Humphreys, *et al.*, under review) |
| * posts\_LIWC.csv | (Pennebaker *et al.*, 2015; Jagfeld, Lobban, Davies, *et al.*, under review) |
| * users.csv | (Jagfeld *et al.*, 2021) |

Additionally, the texts of the 267,576 posts in the BD Subreddit Corpus used in Exploratory Study 1 in Jagfeld, Lobban, Humphreys, et al. (under review) can be requested. The corpus includes 34,995 comments from Reddit users with a self-reported BD diagnosis made in April and May 2019 that are not part of the S-BiDD dataset.

* BD\_Subreddit\_Corpus.txt

posts\_meta.csv

|  |  |
| --- | --- |
| Variable | Description |
| id | Unique id of the post (pseudonymised, does not correspond to real Reddit id) |
| user\_id | Unique id of the post author (pseudonymised, does not correspond to real Reddit id) |
| type | Type of this post: 'c' for comment, 's' for submission |
| created\_at\_UTC | Utc timestamp when post was posted on Reddit provided by Reddit API |
| subreddit\_name | Name of the subreddit in which this post was posted (access subreddit as reddit.com/r/subreddit\_name) provided by Reddit API |
| created\_at | Timestamp of post in users’ local time (based on inferred location) |
| lang | Post language inferred by langid.py (Lui and Baldwin, 2012) |
| text\_wordcount | Number of tokens in this post according to NLTK’s (version 3.4.1) TwitterTokenizer (Bird, Loper and Klein, 2009), assuming English language |
| bd\_diagnosis\_post | Whether this post contains a self-reported BD diagnosis statement (see (Jagfeld *et al.*, 2021)) |
| mentions\_bd | Whether this post contains a BD term (see (Jagfeld *et al.*, 2021)) |
| in\_mh\_subreddit | Whether this post was posted in any MH subreddit (see (Jagfeld, Lobban, Davies, *et al.*, no date)) |
| subreddit\_type | Type of subreddit in which post was posted: bd (Bipolar Disorder (BD)), mh (Mental Health (MH)), other (neither in BD nor MH subreddits list) |

posts\_texts.csv

|  |  |
| --- | --- |
| Variable | Description |
| id | Unique id of the post (pseudonymised, does not correspond to real Reddit id) |
| text | Post text (for submissions contains the title and text body separated by a space character)  Urls in the text were automatically replaced by subURLaddress[[1]](#footnote-1). |

posts\_LIWC.csv contains the same variables as posts\_meta.csv plus the output of Linguistic Inquiry and Word Count (LIWC) tool (Pennebaker *et al.*, 2015) used in Jagfeld et al. (no date).

|  |  |
| --- | --- |
| Variable | Description |
| WC | Number of words in the post according to LIWC, can differ from text\_wordcount |
| posemo | LIWC positive emotion score |
| anx | LIWC anxiety score |
| anger | LIWC anger score |
| sad | LIWC sadness score |
| i | LIWC first person singular pronouns score |

users.csv

|  |  |
| --- | --- |
| Variable | Description |
| id | Unique user id (pseudonymised, does not correspond to real Reddit id) |
| created\_at | Utc timestamp of account creation provided by Reddit API |
| n\_posts | Total number of posts (submissions and comments) of this user in the dataset |
| n\_submissions | Total number of submissions of this user in the dataset |
| n\_comments | Total number of comments of this user in the dataset |
| first\_post\_created\_at | UTC timestamp of first post by this user in the dataset |
| days\_to\_first\_post | Days between account creation and timestamp of first post by this user in the dataset (post.created\_at - user.created\_at) |
| active\_days | Days between first and last post of this user in the dataset |
| activity | n\_posts / active\_days |
| User characteristics derived from post texts + meta data via NLP – see (Jagfeld *et al.*, 2021) | |
| gender | Automatically assigned binary gender identity (masculine/feminine) |
| dob | Automatically assigned date of birth |
| country | Automatically assigned country of origin |
| comorbidities | Other self-reported MH diagnoses of the user |
| avg\_posting\_age | Average age of the user when they made a post in the dataset |
| age\_group | Age group assigned based on avg\_posting\_age (13-17: teenager, 18-29: young-adult, 30-49: adult, 50-64: middle-aged, 65+: senior) |

# REFERENCES

Bird, S., Loper, E. and Klein, E. (2009) *Natural Language Processing with Python*. O’Reilly Media Inc. Available at: https://www.nltk.org/.

Jagfeld, G. *et al.* (2021) ‘Understanding who uses Reddit: Profiling individuals with a self-reported bipolar disorder diagnosis’, in *Proceedings of the Seventh Workshop on Computational Linguistics and Clinical Psychology: Improving Access at NAACL 2021*.

Jagfeld, G., Lobban, F., Humphreys, C., *et al.* (under review) ‘How do people with a bipolar disorder diagnosis talk about personal recovery in peer online support forums? A corpus framework analysis using POETIC’.

Jagfeld, G., Lobban, F., Davies, R., *et al.* (under review) ‘Posting patterns in peer online support forums and their associations with emotions and mood in bipolar disorder: exploratory analysis’.

Lui, M. and Baldwin, T. (2012) ‘langid.py: An Off-the-shelf Language Identification Tool’, *Aclweb.Org*, (July), pp. 25–30. Available at: http://www.aclweb.org/anthology-new/P/P12/P12-3005.pdf.

Pennebaker, J. W. *et al.* (2015) ‘The development and psychometric properties of LIWC2015’, *Austin, TX: University of Texas at Austin*, pp. 1–22. doi: 10.15781/T29G6Z.

1. See https://github.com/glorisonne/reddit\_bd\_mood\_posting\_mh/blob/main/replace\_urls.py [↑](#footnote-ref-1)