Schizophrenia Bulletin
https://doi.org/10.1093/schbul/sbac058

COMMENTARY

Is Collaborative Open Science Possible With Speech Data in Psychiatric Disorders?

Lena Palaniyappan*[,1,2,3], Maria F. Alonso-Sanchez[2,4], and Brian MacWhinney[5]

[1]Department of Psychiatry, Douglas Mental Health University Institute, McGill University, Montreal, QC, Canada; [2]Robarts Research Institute, Western University, London, ON, Canada; [3]Department of Medical Biophysics, Western University, London, ON, Canada; [4]CIDCL, Fonoaudiología, Facultad de Medicina, Universidad de Valparaíso, Valparaíso, Chile; [5]Department of Psychology, Carnegie Mellon University, Pittsburgh, PA, USA

*To whom correspondence should be addressed; Douglas Mental Health University Institute, 6875 Boul. LaSalle, Verdun, Montreal, QC H4H 1R3, Canada; tel: (514)761-6131, fax: (514)888-4064; e-mail: lena.palaniyappan@mcgill.ca

Provision of mental health care is almost entirely built on a singular medium—naturally occurring spoken language conversations. However, datasets of spoken language from patients experiencing mental health issues are surprisingly difficult to obtain. In this commentary, we discuss some of the reasons behind this, and highlight successful approaches adopted in other areas of clinical linguistics and pose some ways forward, especially for the study of psychosis.

Barriers to Sharing Speech Data

Across disciplines, researchers are rapidly adopting Open Science principles for data sharing. This movement encourages researchers, clinicians, and institutions to provide fully open access to research data, programs, and publications. For example, the National Institutes of Health's *Strategic Plan for Data Science* requires that newly funded research projects share data in accord with the FAIR principles[1] for open access and that they include in their budget requests for the resources necessary to complete open access. Although many disciplines, funding agencies, researchers, journals, libraries, and institutions have adopted this new model, the movement has also encountered significant resistance, particularly for open sharing of spoken language data, including spoken language data from clinical populations (SLDCP). We can identify at least 6 barriers to open sharing of SLDCP.[2] Some of these barriers come from the interpretation of regulations by various institutions, while others pertain to the prevailing public perception regarding SLDCP. Here we consider each of these barriers and the ways in which systems, such as TalkBank[3] or Databrary[4] manage to overcome them. With emerging collaborative efforts to study language in psychosis (eg, https://discourseinpsychosis.org/), we anticipate the commentary here to eventually inform "speech bank" infrastructures for psychiatric disorders.

1. *Informed consent.* A frequent objection to the sharing of SLDCP is that it violates participants' rights of privacy and confidentiality. Such usage would be a violation if there had been no informed consent from the participants for sharing of their data—this is, unfortunately, the case for many existing speech samples from clinical populations, precluding retrospective sharing. In these cases, re-contacting participants to obtain consent for data sharing is an option, if consent for such re-contact is in place. In the absence of consent to re-contact, institutional review boards (IRBs) may be able to grant a "waiver," ie, modifying the initial consent parameters (see https://conp.ca/ethics-toolkit/). Some national laws also provide alternatives for re-consenting for scientific purposes.[5] Explicitly stating in informed consent forms that the data will be made available to qualified researchers (holding an identifiable position in an academic or research enterprise wherein research activities are governed by a code of conduct on academic integrity) and that it can be removed from a sharing portal if the participant requests removal, will address this barrier. Qualified researchers can be vetted by an interview process including a signed agreement form by a governance body managing access to the SLDCP (as in the case of HomeBank that stores recordings from children at home settings: https://homebank.talkbank.org/). For SLDCP, there is usually the further stipulation that access requires a password that is only given to researchers and clinicians at established institutions with limits placed on the purpose for which the data is used (eg, academic research, education, noncommercial use). Consenting can also be made dynamic, so permission is in place to "feedback" to the research subjects about the overall use of the data and if a participant changes their mind after a period of time, their contributed data can be deleted from the speech bank.[6]

2. *Deidentification.* Some IRBs and national policies may further require that the data be deidentified, even if there is full informed consent and password control. For audio data, this can be done by avoiding the use of last names and addresses when recording. This requires appropriate prompts and reminders before and during data acquisition (as in DISCOURSE in psychosis protocol). Some IRBs have suggested that spoken language samples could be identified through the use of a "voiceprint." However, without the establishment of a national database of voiceprints, this is not technically possible.[7,8] In fact, the term voiceprint is considered misleading by some, as it gives the impression voice data is equivalent to unique fingerprints, which is not the case.[9] To preclude the application of advanced technologies in the future to the shared data, sharing can be limited to data from constrained speech elicitation tasks rather than using "always-listening" devices. For audio data obtained from speech tasks, screening and manual curation to "bleep-out" personal identifiers can be done with participant input. Third, as voice carries biometric personal information, sharing can be limited to typed transcripts rather than audio files, reducing the risk of inferring the characteristics of the speaker. For video samples, deidentification requires either facial blurring or the replacement of personal images with avatar images (https://getrad.co). However, many IRBs will allow for sharing of password-protected video data, given adequate informed consent. For example, General Data Protection Regulation (GDPR) regulations (European Union) permit sharing of identifiable data for scientific purposes that cannot be fulfilled by deidentified data when there is informed consent while requiring deidentification (pseudonymization) for risk mitigation and to comply with data minimization and storage limitations.
3. *Credit assignment.* Researchers are often worried that competing researchers could use their shared data to scoop them by publishing their results before they have a chance to do so themselves.[10] TalkBank deals with this issue by allowing a period of the embargo on data usage (eg, 2 years), during which the data are included in the database, but not made available on the web. Once the data are made available, researchers can ensure that they receive credit by requiring that any use of corpus data include citation of the corpus (using assigned DOIs or digital object identifiers) and at least one previous publication from the data contributor. This allows for citation crediting through scholar.google.com to judge the impact of a dataset.
4. *Use and misuse.* Researchers often express the fear that their data could be misinterpreted or used in some unethical way. In practice, misuse of this type has never occurred, at least for the databases affiliated with the TalkBank system. However, to avoid possible misunderstandings, sharing could be restricted to vetted qualified researchers who agree to a code of conduct, with intended use proposed and pre-approved by a governing body.
5. *Workload.* For certain types of data, inclusion in a data repository may involve significant work in terms of transcription and data file organization. This type of work can be particularly difficult when the repository requires that data be transcribed in a specific format, as is the case for TalkBank. To lower this barrier, funding agencies provide resources to TalkBank and similar projects to assist researchers and workers in the database to achieve correct data formatting and curation. A positive result of this process is that, once the data are included in the proper TalkBank format, many types of additional analyses and comparisons across datasets become possible through the use of TalkBank tools.
6. *Jurisdictional barriers.* The GDPR regulations of the European Union require that identifiable data collected from European participants not be transferred to other jurisdictions, unless these jurisdictions are pre-approved under an "adequacy decision," have special agreements with the EU, or sign on to the standard contractual clauses of GDPR. Similar restrictions may exist in other jurisdictions. The most straightforward way of dealing with this GDPR restriction is to render the data anonymous (ie, deidentify and remove the means by which singled-out data can be linked to a natural person[11]). A second method would be to establish repositories in countries of the European Union that make data available in a format that matches the requirements of a centralized repository. We can refer to this as a federated content access (FDA) system. Such a configuration provides a greater level of control for contributors and their institutions, but it also requires close adherence to data format standards and systematic installation of the database management system. While individual rights (eg, right to be forgotten) in the wake of scientific data biobanking is an emerging area of debate,[12] successful biobanks (eg, UK BioBank) allow participants to withdraw at any time for any reason.

For SLDCP data from aphasia, apraxia of speech, traumatic brain injury, stuttering, autism spectrum disorder, specific language impairment, and right hemisphere damage, the TalkBank system has managed to overcome all of the above-listed barriers, thereby creating the largest open-access repository for SLDCP. These methods can easily be extended to include data on mental illnesses. For this type of data, however, there are additional barriers that arise from researcher and care-provider perspectives. One

approach to this concern could involve co-designing speech studies with consenting patients and enabling them to interact with their own data and to choose the level of anonymization with which they are comfortable. See Hauglid[13] (in this issue) for other legal and ethical issues that arise from Natural Language Processing (NLP) applications.

The Need for Open Science

Accelerating research with SLDCP requires cross-disciplinary and international collaborations that can fully exploit the unprecedented developments occurring in various domains of clinical linguistics. Cross-language and cross-cultural validations in most areas of SLDCP are scarce, greatly affecting the generalizability of observations. For example, while most patients with psychosis across the globe do not speak in English, studies leveraging NLP are almost exclusively in English. Harmonization (ie, achieving content equivalence) requires several considerations, starting from shared methods and protocols for data acquisition (see Chandler et al, this issue for further discussion). Multiple collaborative efforts that overcome the barriers listed above are essential to interrogate and overcome asymmetries in cultural, social, and geographical factors that are highly relevant for developing NLP applications in mental health.

Rapid open sharing of genetic sequences provided critical support for the scientific efforts against the COVID pandemic.[14] Combating psychiatric disorders with a similar rigor requires a commitment to sharing speech and language data—the most important clinical tool in mental health. It also requires adherence to shared methods for data elicitation and analysis which can then serve as a basis for treatment assessment. Immediate access to speech-based objective measures from consenting patients will make clinical studies more replicable and will open the door to contrasting analyses that target a common dataset. The value of an Open Science ecosystem for SLDCP has been demonstrated in other clinical areas, such as aphasia,[3] dementia,[15] or stuttering[16] with cumulative knowledge on policy frameworks rapidly emerging elsewhere.[17] Given its great promise for understanding and treating psychosis, it is imperative that researchers, clinicians, universities, and funders work together to tear down the barriers to a full implementation of Open Science. We owe it to our patients and their families to make this commitment.

Acknowledgments

We appreciate the members of the Steering Committee of Discourse in Psychosis for several discussions on the material summarized here.

L.P. reports personal fees from Otsuka Canada, SPMM Course Limited, UK, Canadian Psychiatric Association; book royalties from Oxford University Press; investigator-initiated educational grants from Janssen Canada, Sunovion and Otsuka Canada out-side the submitted work. L.P. is the convener of the DISCOURSE in psychosis consortium (www.discourseinpsychosis.org). Brian MacWhinney receives support from NIH grants DC008524 and HD082736. MAS reports no relevant conflicts.

Funding

M.F.A.-S. is supported by the National Agency for Research and Development (ANID), Scholarship Program, Becas Chile 2019, Postdoctoral Fellow 74200048 (MA). The authors acknowledge the support from Tanenbaum Open Science Institute to DISCOURSE in Psychosis (McGill University). L.P. acknowledges personal chair support from the Tanna Schulich Endowment (Schulich School of Medicine and Dentistry, Western University) and Monique H. Bourgeois Endowment (The Douglas Research Centre, McGill University). B.M. acknowledges support from NIH Grant DC1090506.

References

1. Wilkinson MD, Dumontier M, Aalbersberg IJ, et al. The FAIR Guiding Principles for scientific data management and stewardship. *Sci Data.* 2016;3:160018.
2. Houtkoop BL, Chambers C, Macleod M, et al. Data sharing in psychology: a survey on barriers and preconditions. *Adv Methods Pract Psychol Sci.* 2018;1:70–85.
3. MacWhinney B. Understanding spoken language through TalkBank. *Behav Res Methods.* 2019;51:1919–1927.
4. Gilmore RO, Kennedy JL, Adolph KE. Practical solutions for sharing data and materials from psychological research. *Adv Methods Pract Psychol Sci.* 2018;1:121–130.
5. Gefenas E, Lekstutiene J, Lukaseviciene V, et al. Controversies between regulations of research ethics and protection of personal data: informed consent at a cross-road. *Med Health Care Philos.* 2022;25:23–30.
6. Teare HJA, Prictor M, Kaye J. Reflections on dynamic consent in biomedical research: the story so far. *Eur J Hum Genet.* 2021;29:649–656.
7. Yuan J, Liberman M. Speaker identification on the SCOTUS corpus. *J Acoust Soc Am.* 2008;123:3878.
8. Togneri R, Pullella D. An overview of speaker identification: accuracy and robustness issues. *IEEE Circuits Syst Mag.* 2011;11:23–61.
9. Nautsch A, Jiménez A, Treiber A, et al. Preserving privacy in speaker and speech characterisation. *Comput Speech Lang.* 2019;58:441–480.
10. Dorta-González P, González-Betancor SM, Dorta-González MI. To what extent is researchers' data-sharing motivated by formal mechanisms of recognition and credit? *Scientometrics.* 2021;126:2209–2225.

11. Mourby M, Mackey E, Elliot M, et al. Are "pseudonymised" data always personal data? Implications of the GDPR for administrative data research in the UK. *Comput Law Secur Rev.* 2018;34:222–233.
12. Staunton C. Individual rights in biobank research under the GDPR. In: Slokenberga S, Tzortzatou O, Reichel J, eds. *GDPR and Biobanking: Individual Rights, Public Interest and Research Regulation across Europe*. Cham: Springer International Publishing; 2021: 91–104.
13. Hauglid MK. What's that noise? Interpreting algorithmic interpretation of human speech as a legal and ethical challenge. *Schizophr Bull.* 2022;48(5):960–962. doi:10.1093/schbul/sbac008.
14. Besançon L, Peiffer-Smadja N, Segalas C, et al. Open science saves lives: lessons from the COVID-19 pandemic. *BMC Med Res Methodol.* 2021;21:117.
15. Luz S, Haider F, de la Fuente Garcia S, Fromm D, MacWhinney B. Editorial: Alzheimer's dementia recognition through spontaneous speech. *Front. Comput. Sci*. 2021;3:1–4. doi:10.3389/fcomp.2021.780169.
16. Ratner NB, MacWhinney B. Fluency Bank: a new resource for fluency research and practice. *J Fluency Disord.* 2018;56:69–80.
17. Granados Moreno P, Ali-Khan SE, Capps B, et al. Open science precision medicine in Canada: points to consider. *FACETS.* 2019;4:1–19.