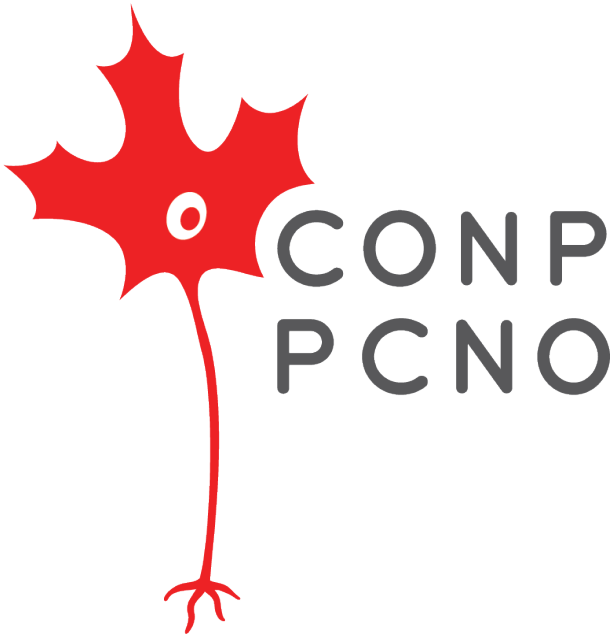
Supported by: Fondation Brain Canada Foundation
with financial support from Health Canada
avec le soutien financier de Santé Canada



Canadian Open Neuroscience Platform
Plateforme Canadienne de Neurosciences Ouvertes

The CONP Privacy and De-identification Toolkit

By: Bartha Maria Knoppers, Michael Beauvais, Ann Cavoukian, John Clarkson, Lindsay Green-Noble, Judy Illes, Jason Karamchandani, Roland Nadler, Dylan Roskams-Edris, and Walter Stewart

Date: February 28, 2022
Version: 1.0.3

CONP Privacy and De-identification Toolkit [1]

The fundamental goal of protecting participant privacy is to prevent the identification of an individual's data by others. In keeping with relevant guidance from Canadian privacy authorities, data must present a low or very low likelihood of individual re-identification before open release, roughly analogous to a maximum threshold of 9% risk of re-identification. Open access datasets should never contain individually identifying information such as names, health card numbers, or social insurance numbers.

Preparing your data: de-identification techniques

The CONP Portal brings together diverse datasets that contain many different types of information in several modalities and formats, e.g., structural and functional MRI, EEG, and behavioural results. The guidance below seeks to help researchers prepare neuroscience data for deposit in Canadian open-access data repositories that accept de-identified data, including the CONP Portal.

This guide is not comprehensive and may need to be tailored to your data. **Researchers bear the responsibility to ensure that tools are applied properly and that information is de-identified before sharing**.

General tools

Tool name	Description	Link(s)
Portage Network's De-Identification Guidance	Guidance about de-identification for many data types and research uses, includes neuroimaging and medical data de-identification information.	https://doi.org/10.5281/zenodo.4270551
NITRC's De-identification Toolbox	Java application that removes identifying information from neuroimaging datasets.	https://www.nitrc.org/projects/de-identification/ Research paper describing the tool
OpenAIRE's Amnesia	Java application that removes identifying information from delimited text files.	https://amnesia.openaire.eu/index.html

Image headers

Tool name	Description	Link
pydicom's deid	Removes information from image headers (customizable).	https://pydicom.github.io/deid/
dicomanon (for MATLAB)	Removes confidential medical information from the DICOM file file_in and creates a new file file_out with the modified values. Image data and other attributes are unmodified.	https://www.mathworks.com/help/images/ref/dicomanon.html

Supported by: Fondation Brain Canada Foundation
with financial support from Health Canada
avec le soutien financier de Santé Canada

Facial and dental features

Tool name	Description	Link
FMRIB Software Library's Brain Extraction Tool (BET)	Removes non-brain tissue from whole-head images.	https://fsl.fmrib.ox.ac.uk/fsl/fslwiki/BET
AFNI's 3dSkullStrip	Extracts brain tissue from MRI T1-weighted images.	https://afni.nimh.nih.gov/pub/dist/doc/program_help/3dSkullStrip.html
Laboratory for Computational Neuroimaging's FreeSurfer Software Suite	Comprehensive software suite that includes tools for skull stripping.	https://surfer.nmr.mgh.harvard.edu/
Peer Herholz's BIDSonym	Gathers T1w images from a BIDS dataset and applies a selected de-identification algorithm. Either: ▪ mri_deface ▪ PyDeface ▪ Quickshear ▪ mridefacer	https://github.com/PeerHerholz/BIDSonym

Cf also: https://community.imagingqa.com/docs

Synthetic data

Synthetic data are data that have been generated from either "real" data or models and that possess the same statistical properties as the original data. While not completely free of re-identification risks, synthetic data are increasingly popular for machine-learning applications.

Tool name	Description	Link
SYLLS' synthpop package for R	Creates synthetic versions of data.	https://cran.r-project.org/web/packages/synthpop/index.html Research paper describing the tool

Articles

Vaden, Kenneth I., Mulugeta Gebregziabher, Dyslexia Data Consortium, and Mark A. Eckert. "Fully Synthetic Neuroimaging Data for Replication and Exploration." *NeuroImage* 223 (December 1, 2020): 117284. https://doi.org/10.1016/j.neuroimage.2020.117284.

El Emam, Khaled, Lucy Mosquera, and Jason Bass. "Evaluating Identity Disclosure Risk in Fully Synthetic Health Data: Model Development and Validation." *Journal of Medical Internet Research* 22, no. 11 (November 16, 2020): e23139. https://doi.org/10.2196/23139.

Deciding between open, registered, and controlled access

Access model	Description	Identifiability of data
Open	Accessible with minimal restrictions or verifications.	Fully de-identified data (both direct and indirect identifiers removed) and/or aggregate data.
Registered	Accessible only to users who have an account that has identified them as a *bona fide* researcher.	De-identified data and/or aggregate data where inferences may be made about indirectly identifying individual records.
Controlled	Accessible only upon review of a data access application, which includes prior approval by a research ethics board.	Individual-level data with direct identifiers removed or replaced by a code.

Articles

Dyke, Stephanie O. M., Mikael Linden, Ilkka Lappalainen, Jordi Rambla De Argila, Knox Carey, David Lloyd, J. Dylan Spalding, et al. "Registered Access: Authorizing Data Access." *European Journal of Human Genetics* 26, no. 12 (December 2018): 1721–31. https://doi.org/10.1038/s41431-018-0219-y.

Dyke, Stephanie O. M., Emily Kirby, Mahsa Shabani, Adrian Thorogood, Kazuto Kato, and Bartha M. Knoppers. "Registered Access: A 'Triple-A' Approach." *European Journal of Human Genetics* 24, no. 12 (December 2016): 1676–80. https://doi.org/10.1038/ejhg.2016.115.

Additional resources

Basic concepts

Chapter 5: Privacy and Confidentiality of the *Tri-Council Policy Statement: Ethical Conduct for Research Involving Humans* https://ethics.gc.ca/eng/tcps2-eptc2_2018_chapter5-chapitre5.html.

Sensitive Data Expert Group. "Sensitive Data Toolkit for Researchers Part 1: Glossary of Terms for Sensitive Data Used for Research Purposes," September 30, 2020. https://doi.org/10.5281/zenodo.4088946.

Sensitive Data Expert Group. "Sensitive Data Toolkit for Researchers Part 2: Human Participant Research Data Risk Matrix," October 1, 2020. https://doi.org/10.5281/zenodo.4088954.

Canadian Institutes of Health Research. "CIHR Best Practices for Protecting Privacy in Health Research," September 15, 2005. https://cihr-irsc.gc.ca/e/documents/et_pbp_nov05_sept2005_e.pdf.

Beauvais, Michael J.S., Bartha Maria Knoppers, and Judy Illes. "A Marathon, Not a Sprint – Neuroimaging, Open Science and Ethics." *NeuroImage* 236 (August 1, 2021): 118041. https://doi.org/10.1016/j.neuroimage.2021.118041.

Data management plans

Morissette, Erica, Lina Harper, Isabella Peters, Felicity Tayler, and Stefanie Haustein. "Data Management Plan Template: Open Science Workflows," April 9, 2021. https://doi.org/10.5281/zenodo.4701021.

Strauss, Ted. "Data Management Plan Template: Neuroimaging in the Neurosciences," April 9, 2021. https://doi.org/10.5281/zenodo.4673558.

Creation of open access datasets

Tremblay-Mercier, Jennifer, Cécile Madjar, Samir Das, Alexa Pichet Binette, Stephanie O. M. Dyke, Pierre Étienne, Marie-Elyse Lafaille-Magnan, et al. "Open Science Datasets from PREVENT-AD, a Longitudinal Cohort of Pre-Symptomatic Alzheimer's Disease." BioRxiv, November 30, 2020, 2020.03.04.976670. https://doi.org/10.1101/2020.03.04.976670.

Tremblay-Mercier, Jennifer, Cécile Madjar, Samir Das, Stephanie O. M. Dyke, Pierre Étienne, Marie-Elyse Lafaille-Magnan, Pierre Bellec, et al. "Creation of an Open Science Dataset from PREVENT-AD, a Longitudinal Cohort Study of Pre-Symptomatic Alzheimer's Disease." BioRxiv, March 5, 2020, 2020.03.04.976670. https://doi.org/10.1101/2020.03.04.976670.

Data governance

Eke, Damian, Amy Bernard, Jan G. Bjaalie, Ricardo Chavarriaga, Takashi Hanakawa, Anthony Hannan, Sean Hill, et al. "International Data Governance for Neuroscience." PsyArXiv, June 1, 2021. https://doi.org/10.31234/osf.io/esz9b.

[1] Developed by the Ethics and Governance Committee of the Canadian Open Neuroscience Platform. Members: Bartha Maria Knoppers (Chair), Michael Beauvais (Manager), Ann Cavoukian, John Clarkson, Lindsay Green-Noble, Judy Iles, Jason Karamchandani, Roland Nadler, Dylan Roskams-Edris, and Walter Stewart.

Supported by: Fondation Brain Canada Foundation
with financial support from Health Canada
avec le soutien financier de Santé Canada