# Plan Overview

*A Data Management Plan created using DMPonline*

**Title:** Creating a research repository to assess the validity of clinical assessments of autism symptomatology and traits in genetic syndromes

**Creator:**Natali Bozhilova

**Affiliation:** University of Surrey

**Template:** DCC Template

**Project abstract:**

Rationale: We and others have shown that genetic syndromes give rise to distinct patterns of autism (ASD) characteristics, making assessment and diagnosis of ASD more complex in these populations. This level of complexity is likely to reduce the sensitivity and specificity of existing ASD diagnostic assessments in these populations, explaining the lower recognition rates. Our pilot study has thus showed a low concordance rate between traditional ASD diagnostic assessments in 240 individuals with genetic syndromes compared to individuals with idiopathic ASD. Aims: The primary aim of the study is to assess the validity, sensitivity and specificity of ASD diagnostic assessments across syndrome groups in a larger and multinational sample. Method: This study will combine our existing datasets with datasets from national and international collaborators. The main focus will be on a) gold standard ASD diagnostic assessments as part of research protocols, or clinical trials/visits, b) formal clinical report/confirmation as part of the ASD diagnostic assessment. The data will be stored in a shared access repository (e.g., RedCap), which will facilitate collaboration across multiple sites/research groups. Collaboration agreements will allow data sharing with all the collaborators. Analyses: Convergent validity between ASD diagnostic assessments (i.e., ADOS, SCQ and ADI-R) will be assessed with correlational analyses (including polyserial correlations to understand the correlation between quantitative and ordinal variables). Construct validity of ASD diagnostic assessments will be assessed with one-way ANOVA and mixed models to understand the differences between syndrome groups and individuals with idiopathic ASD. Receiver operating characteristics (ROC) will be used to examine diagnostic accuracy (i.e., specificity and sensitivity) and the optimal cut-off point for these measures in the syndrome groups. Outcome: The primary outcome is to understand the accuracy of existing ASD assessments across a very large number of syndrome groups. Our findings will provide evidence for the need to develop measures, which will ensure greater precision in assessment and diagnosis of ASD in these groups.

**ID:** 80556

**Start date:** 06-09-2021

**End date:** 30-08-2023

**Last modified:** 02-03-2022

as they would like in their own plans, and customise it as necessary. You do not need to credit the creator(s) as the source of the language used, but using any of the plan's text does not imply that the creator(s) endorse, or have any relationship to, your project or proposal

# Creating a research repository to assess the validity of clinical assessments of autism symptomatology and traits in genetic syndromes

## Data Collection

**What data will you collect or create?**

**1.1 Type of study:** The study aims to combine pre-existing data from gold standard assessments of autism symptomatology in genetic syndromes into a fully anonymised database.

**1.2 Types of data:** *Quantitative data*: Clinical assessments of autism symptomatology: gold standard assessments of autism symptomatology (e.g., social communication questionnaire (SCQ), Autism Diagnostic Observation Schedule (ADOS)). The data will include raw and standardised scores. *Demographics questionnaires*: demographics, socioeconomic status, home environment, record of educational / clinical support and clinical/genetic diagnosis.

**1.3 Format and scale of the data:** Our and collaborators' data are currently formatted as .csv, .spss, .sav and .xlsx files. Clinical face-to-face assessments will be in a .pdf, .csv and .txt format. These formats are standardised, widely used within both research and clinical settings and easily interchangeable, which ensures the long-term usability of data. The data will be compiled in a secure server, which will be hosted at REDCap supported by university servers (https://projectredcap.org/) to facilitate partnership with the collaborators. All the data formats and storage software (REDCap) allow sharing and long-term access to the data. The volume/size of all files will not exceed 10-20 GB (10 GB of raw data, 5-10 GB processed data). The volume/size of individual databases from each of the collaborators will not exceed 20 MB, which will allow efficient data sharing.

**How will the data be collected or created?**

**2.1 Data collection:** The data will be collated and deposited following GDPR requirements by Dr Joanna Moss (PI) and research fellow/data manager (Dr Natali Bozhilova). The collaborators will be required to fully anonymise their data prior to data submission to the PI and the data manager. Following anonymisation and prior to receipt, the collaborators will submit the data via the secure server/REDCap (Surrey drop-off service). Collaboration agreements will also be signed. Upon receipt, each participant's data will receive a new unique ID code for the purposes of our project. The original ID number/code will be deleted.

**2.2 Data quality and data sharing:** Data will be stored using REDCap's secure web application. Data sharing agreements between the PI and each of the collaborators will be in place to ensure data protection during data sharing, data management/storage and data analysis. Each dataset will be checked for completeness and overall quality by the data manager following a study-specific protocol (i.e., less than 30% of missing data and presence of key autism-related assessments). As part of the data sharing, the collaborators will also have to share the reference number of their ethical approval(s) for the initial data collection.

**2.3 Managing, storing and curating data:** The PI and the data manager will build a brand new repository with each of the datasets. The repository will be stored on REDCap and accessible via REDCap. Repository access will be available upon request (i.e., e-mail owners) at the time of publication and as part of a close collaboration between the study investigators and the researcher, who requests access to the data.

As part of the process, the data manager will adopt the same naming conventions, folder organisation and version control for each of the files from the collaborators. A detailed description of each of the variables, files and folder structure will be stored as a ReadMe .txt file alongside the data. This .txt file will help existing and future data managers/research staff to easily access, understand and analyse the data the newly built repository, thereby making it possible to replicate the findings and reuse the data for new purposes/projects.

## Documentation and Metadata

**What documentation and metadata will accompany the data?**

**3.1 Metadata standards and data documentation:** In addition to the ReadMe.text file for each dataset, we will provide a description of how/when/where the data have been collected/collated (i.e., a copy of study protocols when available). This information will, along with field codes, will be stored in a ReadMe .txt file together with the newly built repository and the original data on the server and REDCap during the duration of the project and long-term archiving. Both collaborators and research team will have access to the ReadMe .txt file. Quantitative data will be stored with unique ID research codes, which will not link to personal or

identifying data. Separate from the repository, a codebook (employing Data Documentation Initiative [DDI] standard), indicating the question, potential responses (or range) and type of data will be produced to enable interpretation of variables and label names. This information will be available upon request.

**3.2 Data preservation strategy and standards:** Data from this study will be stored for a minimum of 10 years. Long-term data storage will be in .csv files using a standard (simple) ASCII coding scheme in line with UK Data Service recommendations (https://www.ukdataservice.ac.uk/manage-data/format/recommended-formats).

## Ethics and Legal Compliance

**How will you manage any ethical issues?**

**4.1 Identity protection:** Prior to data sharing, data will be fully anonymised to remove personal information according to GDPR regulations and data sharing/collaborative agreements will be signed between the PI and all collaborators. Statements regarding the prevention of re-identification of participants will be included in the data access policy, drawn from similar examples (e.g. Imagine-ID, UK Biobank). To avoid linkages, brand new participant IDs will be used for each data sharing project. Shared data will not include variables that risk participant identification (e.g. post code, national insurance, date of birth). Although the syndromes are rare, large samples minimise the risk of identifying specific individuals. In line with Imagine-ID data access policy, where identification is a risk (https://imagine-id.org/wp-content/uploads/2019/04/Data-Access-Policy-16.04.19-1.pdf), data recipients will sign that they will make no attempt to identify any research participants (i.e., non-disclosure agreements).

**4.2 Governance review and ethical approval:** The project will commence following governance and ethics review and approval granted by the University of Surrey. Each collaborator will agree to submit the reference number, or a letter of favourable opinion from the relevant ethics committee alongside with their dataset(s). Some of the data might have been collected as part of clinical assessments. In the event of such instances, the collaborators will have to fully anonymise the data and provide evidence for data use in future/our project, which each participant would have read and signed prior to data sharing. Surrey legal team will advise regarding all legacy data sharing aligns with existing participant consent (all consented to sharing with bona fide researchers). We will seek retrospective consent on rare occasions that this is required; where this is not possible, we will adhere to point 3 of Utrecht University Data Management guide (already approved by Surrey legal team). “Sensitive data can legally be shared without explicit consent if the information given to participants prior to their consent for data collection indicated future use of the data, or if ALL of the following are applicable: 1. The opportunity to gain consent no longer exists or is not practical; 2. The data have been de-identified; 3. There is no risk that publishing or sharing the data will cause harm or contribute to discrimination towards the research participants or subjects; 4. Information sheets and consent forms from the original data collection did not preclude sharing” Point 3 Informed consent for data sharing | Universiteit Utrecht (uu.nl). The process of data sharing will commence following evidence of successful ethical approval or provision of consent forms.

**How will you manage copyright and Intellectual Property Rights (IPR) issues?**

**5.1 Management of Copyright and IPR:** The built-up repository will be owned by PI, which will be detailed in the collaboration agreements. Each collaborator will be allowed access to the built-up repository once the repository is built up.

**5.2 Suitability for sharing:** The repository will be suitable for sharing with bona fide researchers (see MRC definition of bona-fide research MRC-0208212-MRC-policy-and-guidance-on-sharing-of-research-data-from-population-and-patient-studies-Word-version-v01.02.pdf (ukri.org))and will be available at the point of paper publication, after locking the analytical database in collaboration with RCUK, UK Data Service, and Open Science Framework. The option of sharing the data at the point of dissemination might also require to license the repository. In that instance, we will use the CC BY-NC-SA license (https://creativecommons.org/licenses/), which prevents commercialisation of the data and requires any further uses to be shared on the same terms (i.e. non-commercial).

**5.3 Governance of access:** Access to the repository will follow a closer collaboration between the study investigators and the researcher, who might request access to the repository. There will be therefore no formal approval process apart from verification of a status as a bona fide researcher. The guiding principle is that data are a public good and require stewardship that maximises their potential to benefit the public. This is balanced with the need to protect study participants, achieved by defining access ‘levels’ that require progressively closer collaboration with the study investigators to access data with details that could potentially identify participants (considering the rarity of genetic syndromes). Access will not be granted to researchers that fail to meet data access policy terms. A summary of the data use will be published as part of the main paper for transparency and to avoid duplication of access requests.

**5.4 The study team’s exclusive use of the data:** The data will be used exclusively by the project team prior to paper publication and locking of the analytical database. Following this process, the data will become available to other bona fide researchers, who plan to establish a close collaboration with the study investigators at the point of findings dissemination.

## Storage and Backup

**How will the data be stored and backed up during the research?**

**6.1 Data storage:** REDCap , hosted on university servers, will provide sufficient storage space for the data (exceeding 20 GB). We already have a designated lab folder on the server.
**6.2 Data back-up:** The data will automatically backed up (twice a day) in multiple physical locations, hosted at the University of Surrey. The data will be recovered with the support of the cyber security IT team drawing from the automatic back-up reserve. The data will not be stored on laptops, computer hard drives or external storage devices.
**6.3 Formal information/data security standards:** This study will adhere to the highest international information/data security standards following the UK government backed-up scheme, *Cyber Essentials (https://www.ncsc.gov.uk/cyberessentials/overview)* and the regulations imposed by the Information Security Policy of the University of Surrey, in full conformity with GDPR.

**How will you manage access and security?**

**6.4 Main risks to data security:** Database/Repository security will be scrutinised by Surrey IT cyber security team, who will support risk mitigation planning during the project. Risk of unauthorised access to the project space on REDCap will be addressed via encrypted laptops, and password-protected files. Risk of loss of data (damage/loss/theft) will be addressed by ensuring the data is backed up regularly, stored in secure locations, never kept on unencrypted devices or stored on laptops or flash drives. All collaborators and research staff will have secure/encrypted and guided access to the project folder. All existing data are maintained on encrypted servers and will be transferred only via secure protocols. The identity of all participants and ID link to data will be kept on the password protected university computer system at the relevant institutions. We or the rest of the collaborators will receive data that contain anonymised individual identifiers and will not have access to personal identifiers. The PI and data manager will manage the control access to the database/data (as specified in 5.3).
**6.5 Data security after the project duration:** After the project duration, new data will be inputted online with secure web authentication, data logging and Secure Sockets Layer (SSL) encryption by the data manager. The Surrey University server is regularly backed up in multiple locations, is resilient and hardened against hacking and other forms of attacks. The server is located in a secured server room with controlled physical access. All data traffic is encrypted, and an installed network firewall provides secure network access. REDCap system also provides inherent application level security.

## Selection and Preservation

**Which data are of long-term value and should be retained, shared, and/or preserved?**

**7.1 Data preservation:** The entire database will be preserved and archived using UK Data Service, Figshare, Zenodo or Surrey Data Repository Services (https://www.surrey.ac.uk/library/open-research/research-data-management-and-sharing).

**What is the long-term preservation plan for the dataset?**

**7.2 Long-term preservation:** A timestamped static database will be deposited with UK Data Service, Figshare, Zenodo or Surrey Repository Services at the end of the grant period, and periodically thereafter every 2 years. Data users (apart from the PI and the data managers) will not be able to make any changes to the lrepository. REDCap allows this level of flexibility (i.e., different levels of access). The data sharing/ preservation will be done within the remit of the project grant. No additional costs are envisaged once the repository has been built and developed within the scope of the grant.

## Data Sharing

**How will you share the data?**

**8.1 Data sharing:** We will add the study/project/data description to the main publication. After the completion of the project, the data will be shared/available via REDCap upon request (see 5.3).
**8.2 Regulation of responsibilities of users:** The PI and the data manager will ensure the regulation of responsibilities of users.

REDCap is a particularly useful platform because the PI/data manager will be able to regulate the level of access to the data at all times. During the lifetime of the project, the research staff will have secure access to a folder on a server, provided by the University of Surrey. This folder will be used for data capture and initial storage and accessed from a secure, password-protected university computer. Once all existing data has been exported into REDCap, data files will be stored in the secure server folder for the duration of the project to enable team access

**8.3 Discovery by potential users of the research data:**We will also make discovery metadata (which describes the data) available via the main publication paper. The data in the repository will have a persistent data identifier ,DOI, which will be used in all outputs (publications, conferences, social media) from the database. We will also include a data access statement (and DOI) in all publications; the database (and DOI) will also be listed in University of Surrey's data catalogue.

**Are any restrictions on data sharing required?**

**8.4 Restrictions or delays to sharing, with planned actions to limit such restrictions:**
The data will be released on time (i.e., at the time of findings dissemination). Restrictions to data access to bona-fide researchers, who plan to establish a close collaboration with the study investigators (including signing data sharing agreements) (see 5.3. ). The main reasons for these restrictions is to allow preparation time for data release, submission of relevant publications and data protection from misuse.

## Responsibilities and Resources

**Who will be responsible for data management?**

**9.1 Responsibility:** The data manager together with the data management team at Surrey will be responsible for the study-wide management of the data, for ensuring that the data adheres to our agreed file naming conventions, folder organization and version control, and for the creation and management of the metadata. University of Surrey IT services will provide password data storage and transfer and encryption for laptops. Each collaborator will be responsible for ensuring security of the data they provide on their own university system. Oversight of the DMP implementation will be provided by the study PI (Dr Joanna Moss) and data manager (Dr Natali Bozhilova).

**What resources will you require to deliver your plan?**

**9.2 Resources:** The study will require the installation of the REDCap platform on the university servers. No other training or repository charges are expected. The REDCap and UK Data Archive are free of charge, but require formal applications and IT support (REDCap). .