RESEARCH DATA MANAGEMENT PLAN

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| PROJECT | | | | | |
| Title | Statistical evaluation of maternal metabolome data: are positive correlations with SGA in the dataset justified? | | | | |
| Description | Metabolomics is an emerging field that routinely deals with the collection and analysis of very large datasets. In the course of obtaining such data, researchers must be especially careful to monitor and correct for sources of systematic error. This project will use statistical methods, implemented with the R software, to examine an existing set of metabolomic data and evaluate the ability of multivariate statistics to correct for sources of error. | | | | |
| Field of Research | Medical Science; Metabolomics; Biostatistics | | | | |
| DMP created | | Last updated | Project start | | Project end |
| 2016/07/28, 12:28 | | 2016/08/22 | 2016/07/16 | | 2016/10/28 |
| PROJECT CONTRIBUTORS | | | | | |
| Role | Name | Affiliation | Email | Username | ORCiD ([*i*](http://www.library.auckland.ac.nz/services/research-support/orcid)) |
| Student/Project Lead | Kieran Deane-Alder | FMHS, SBS | kdea020@aucklanduni.ac.nz | kdea020 | orcid.org/0000-0003-2762-2240 |
| Data contact | “” | “” | “” | “” | “” |
| Supervisor | Dr. Elizabeth J. McKenzie | FMHS, Liggins Institute | liz.mckenzie@auckland.ac.nz | - | - |
| Co-supervisor (statistics) | Dr. Beatrix Jones | Massey University; Liggins Institute | beatrix.jones@gmail.com | - | - |

|  |  |  |
| --- | --- | --- |
| POLICIES & GUIDANCE | | |
| Related policies | This project will be conducted in accordance with the Researcher Code of Conduct of the University of Auckland, which may be found at:  <https://www.auckland.ac.nz/en/about/the-university/how-university-works/policy-and-administration/research/conduct/code-of-conduct-policy.html> | |
| FUNDING | | |
| Funding agency | Vice Chancellor’s Strategic Development Fund (University of Auckland, 2015) | |
| Funding ID | - | |
| Research Office ID | - | |
| ETHICS & PRIVACY | | |
| Ethics requirements | The data were obtained in 2014 from a study involving humans, following ethics guidelines and approval from the University of Auckland Human Participants Ethics Committee (hereafter UAHPEC). As such, no further ethics approval is required for this project, which merely analyses an existing published dataset.  In addition to this, the data have been anonymised and must remain so under the original requirements with which the data were collected. | |
| How will you manage any ethical issues? | If any part of the data are to be published as part of this analysis, they must remain anonymised. This extends to identifying specific compounds, so the names of analytes must also be anonymised. However, no component of the data is likely to be included in the final report as the software produced will be data agnostic.  Spoken consent has been granted for the data to be stored digitally for the purposes of this project in Google Drive and on Github, as needed, as it is appropriately anonymised. If necessary, a formal agreement can be arranged. | |
| Are there other privacy and/or security requirements? | Not at this time. | |
| DATA ORGANISATION | | |
| **Data collection/ creation** | | |
| What data will you create/ collect? | Data have been collected by the Liggins Institute and extracted from GC-MS spectral output. Data have since been exported into .XLSX and .CSV formats, for use with Microsoft Excel, and the R scripting language, respectively.  The raw data exist as tables in .CSV (comma-separated values) format, for use with the R scripting language. This is a standardised format that is accepted by most software used in metabolomics data analysis and is likely to be highly compatible, on any current operating system, for the foreseeable future.  There are only two data files used by this project, one describing the calculated relative intensities of compounds for each sample; the other is a machine log with injection order times corresponding to when each sample was run. As the files are small in size (<100 KB), they are easily transferable and storable by any means (internet or offline storage). | |
| How will the data be collected/ created? | The data have already been collected and/or created. | |
| What non-digital data/assets will you create/ collect? | None. | |
| **File management** | | |
| How will the data be organised? | Data will be stored on both a laptop computer (operating under Ubuntu 16.04 LTS) and a desktop PC (Microsoft Windows 10). Data is stored offline in a primary folder, ‘/MedSci736/metabolomics-project’, on the computer desktop. Version control is managed by GitHub, in an identically-named repository under the user kdea020, and commits and updates will be stored there regularly.  Obtained values from the experiment are stored in the file “Hair.csv”, whereas the injection order times as “injection\_order\_SGA\_hair.csv”. The project, consisting of an R script to interpret and analyse this data, will be named “metabolomics-project-[version].R”. | |
| **Storage locations** | | |
| How will the data be stored and backed up during the research? | Storage needs are minimal (<100MB). Data will be stored both “in the cloud” and on the laptop and desktop PCs as previously described.  In the event of an incident, redundant backups will be used to restore the work (including in-progress R scripts). In totality, these will include one google drive backup, one backup on GitHub (with full history of revisions) and a physical backup on an external hard drive under the ownership of K. Deane-Alder. K. Deane-Alder will also be responsible for backup and recovery.  Backups will be made whenever the data or R scripts are updated, as needed, or at least weekly – whichever is more frequent. | |
| METADATA & DOCUMENTATION | | |
| What documentation and metadata will accompany the data to support its discovery, (re)use and increase impact? | A “readme.md” file will accompany the git repository and will include a full description of the data, and plain-language instructions for using the R script produced to analyse it. Additionally, the R script itself will be well-commented and offer a simple method for user input (in English only).  This readme.md file will also include a brief explanation of the rationale behind the interpretive statistics used in the project, as well as an explanation of the format and contents in the two aforementioned raw data .CSV files.  If the project’s scope expands such that more detailed metadata is deemed necessary, an .XML will be included with all distributions (public or private) of the software and dataset, following the Dublin Core Metadata Element Set standard (ISO 15836). | |
| Spatial extent | N/A | |
| Temporal extent | N/A | |
| Links | TBD | |
| OWNERSHIP, COPYRIGHT & IP | | |
| **The copyright and other IP is owned/held by:** | | Yes or leave blank |
| The University of Auckland (normal situation for research undertaken by university staff) | | Yes |
| The student (research by research student in the normal course of study, which does not fall into any of the other categories.) | | Yes |
| Joint ownership (research conducted in collaboration: copyright and IP ownership are documented in an agreement between the organisations) | | \* |
| Third party data (data owned by third party or generated under UniServices agreements. | | \* |
| If ownership *is* jointly held, third party or generated under UniServices contract. | [State the relationships, agreements and relative rights to use, store, publish and re-use the data.] | |
| ACTIVE DATA - SHARING & ACCESS CONTROL | | |
| Access to the data during the project will be: | Open (anonymised data and R script). | |
| How will you manage access and security? | The project will be viewable, in an unrestricted fashion, on GitHub. However, the ability to edit files, add files, or in any other way alter the repository or its contents will be restricted to project members. | |
| RETENTION & DISPOSAL | | |
| **Data must be retained after submission of thesis or publication of results for a minimum of:** | | (select) |
| 6 years (standard minimum retention after last publication based on data) | | \/ |
| 10 years (for medical research involving clinical trials from the end of the trial) | | - |
| Until patient reaches 26 years of age, and at least 10 after last treatment (for clinical research involving children) | | - |
| 21 years from the date of filing a patent related to this research | | - |
| Other specified time | | - |
| Details of other time | - | |
| Based on the above, data must be kept until at least | | 2020/10/28 |
| DATA PUBLISHING AND DISCOVERY | | |
| Licencing | Raw data will not be made publically accessible at this stage. Anonymised data, and the R script produced to analyse it, will be available for open view via the project’s git repository (kdea020/metabolomics-project) under the “default” GitHub license for the time being. | |
| Outline how data will be prepared and where it will be published. | The data are not intended to be published at this stage. | |
| LONG-TERM ARCHIVE / PRESERVATION (20+years, if applicable) | | |
| What is the long-term preservation plan for the dataset? | The .csv format of the raw data should remain compatible with future software due to its widespread use and simple format that is easily machine-readable. At this time, long-term preservation is not considered necessary for this dataset. | |

|  |  |
| --- | --- |
| RDM/DMP RESPONSIBILITIES & RESOURCES | |
| Who will be responsible for data management? | Dr. Elizabeth J. McKenzie: data capture.  Kieran Deane-Alder: metadata production, data quality, storage and backup, data archiving & data sharing. |
| What resources will you require to deliver your plan? | A laptop computer (provided by the university) and a desktop computer (under the ownership of K. Deane-Alder) |

References and thanks to:

DCC. (2013). Checklist for a Data Management Plan. v.4.0. Edinburgh: Digital Curation Centre. Available online: http://www.dcc.ac.uk/resources/data-management-plans