# FWO DMP Template - Flemish Standard Data Management Plan

Project supervisors (from application round 2018 onwards) and fellows (from application round 2020 onwards) will, upon being awarded their project or fellowship, be invited to develop their answers to the data management related questions into a DMP. The FWO expects a **completed DMP no later than 6 months after the official start date** of the project or fellowship. The DMP should not be submitted to FWO but to the research co-ordination office of the host institute; FWO may request the DMP in a random check.

At the end of the project, the **final version of the DMP** has to be added to the final report of the project; this should be submitted to FWO by the supervisor-spokesperson through FWO's e-portal. This DMP may of course have been updated since its first version. The DMP is an element in the final evaluation of the project by the relevant expert panel. Both the DMP submitted within the first 6 months after the start date and the final DMP may use this template.

The DMP template used by the Research Foundation Flanders (FWO) corresponds with the Flemish Standard Data Management Plan. This Flemish Standard DMP was developed by the Flemish Research Data Network (FRDN) Task Force DMP which comprises representatives of all Flemish funders and research institutions. This is a standardized DMP template based on the previous FWO template that contains the core requirements for data management planning. To increase understanding and facilitate completion of the DMP, a standardized **glossary** of definitions and abbreviations is available via the following link.

| 1. General Project Information | |
|---|---|
| Name Grant Holder & ORCID | **Maarten De Vos** http://orcid.org/0000-0002-3482-5145 |
| Contributor name(s) (+ ORCID) & roles | **Helene Huts, PhD student (https://orcid.org/0000-0002-0617-1316)** |
| Project number[1] & title | G0C9623N, **ARTIFICIAL INTELLIGENCE (AI) FOR DATA-DRIVEN PERSONALISED MEDICINE** |
| Funder(s) GrantID[2] | D-2023-2247 |
| Affiliation(s) | X KU Leuven<br>☐ Universiteit Antwerpen<br>☐ Universiteit Gent<br>☐ Universiteit Hasselt<br>☐ Vrije Universiteit Brussel<br>☐ Other:<br>Provide ROR[3] identifier when possible: |
| Please provide a short project description | Novel AI strategies will be developed for assessing personalised treatment effects based on retrospective data analysis of randomised clinical trials and interactive explainable tools will be investigated to improve model performance. |

[1] "Project number" refers to the institutional project number. This question is optional since not every institution has an internal project number different from the GrantID. Applicants can only provide one project number.

[2] Funder(s) GrantID refers to the number of the DMP at the funder(s), here one can specify multiple GrantIDs if multiple funding sources were used.

[3] Research Organization Registry Community. https://ror.org/

## 2. Research Data Summary

List and describe all datasets or research materials that you plan to generate/collect or reuse during your research project. For each dataset or data type (observational, experimental etc.), provide a short name & description (sufficient for yourself to know what data it is about), indicate whether the data are newly generated/collected or reused, digital or physical, also indicate the type of the data (the kind of content), its technical format (file extension), and an estimate of the upper limit of the volume of the data[4].

| | | | | *Only for digital data* | *Only for digital data* | *Only for digital data* | *Only for physical data* |
|---|---|---|---|---|---|---|---|
| Dataset Name | Description | New or Reused | Digital or Physical | Digital Data Type | Digital Data Format | Digital Data Volume (MB, GB, TB) | Physical Volume |
| | | ☐ Generate new data<br>☐ Reuse existing data | ☐ Digital<br>☐ Physical | ☐ Observational<br>☐ Experimental<br>☐ Compiled/ aggregated data<br>☐ Simulation data<br>☐ Software<br>☐ Other<br>☐ NA | ☐ .por<br>☐ .xml<br>☐ .tab<br>☐ .csv<br>☐ .pdf<br>☐ .txt<br>☐ .rtf<br>☐ .dwg<br>☐ .tab<br>☐ .gml<br>☐ other:<br>☐ NA | ☐ < 100 MB<br>☐ < 1 GB<br>☐ < 100 GB<br>☐ < 1 TB<br>☐ < 5 TB<br>☐ < 10 TB<br>☐ < 50 TB<br>☐ > 50 TB<br>☐ NA | |
| IMPACT | Dataset and dataset documentation from the IMPACT trial | Reuse existing data | Digital | Experimental | .csv,<br>.pdf | < 100 GB | |
| PIONEER | Dataset and | Reuse existing data | Digital | Experimental | .csv, | < 1 GB | |

[4] Add rows for each dataset you want to describe.

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| | dataset documentation from the PIONEER trial | | | | .pdf,<br>other: .xpt, .xlsx | | |
| ISAACC | Dataset and dataset documentation from the ISAACC trial | Reuse existing data | Digital | Experimental | (yet to be determined) | (yet to be determined) | |
| BeNeBio | Dataset and dataset documentation from the BeNeBio trial | Reuse existing data | Digital | Experimental | (yet to be determined) | (yet to be determined) | |
| MACRO | Dataset and dataset documentation from the MACRO trial | Reuse existing data | Digital | Experimental | .csv,<br>.pdf,<br>other: .sas7bdat | < 100 MB | |
| COLUMBUS | Dataset and dataset documentation from the COLUMBUS trial | Reuse existing data | Digital | Experimental | .csv,<br>.pdf,<br>other: .sav, .xlsx | < 100 MB | |
| BACE | Dataset and dataset documentation | Reuse existing data | Digital | Experimental | .csv,<br>other: .xlsx | < 100 MB | |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| | from the BACE trial | | | | | | |
| Project code and generated data – IMPACT | Preprocessing, model code, evaluation code, model results | Generate new data | Digital | Software, other | .csv, .pdf, other: .tiff, .png, .xlsx, .eps, .py, .ipynb, .R, .Rmd | < 100 MB | |
| Project code and generated data – PIONEER | Preprocessing, model code, evaluation code, model results | Generate new data | Digital | Software, other | .csv, .pdf, other: .tiff, .png, .xlsx, .eps, .py, .ipynb, .R, .Rmd | < 100 MB | |
| Project code and generated data – ISAACC | Preprocessing, model code, evaluation code, model results | Generate new data | Digital | Software, other | .csv, .pdf, other: .tiff, .png, .xlsx, .eps, .py, .ipynb, .R, .Rmd | < 100 MB | |
| Project code and generated data – BeNeBio | Preprocessing, model code, evaluation code, model results | Generate new data | Digital | Software, other | .csv, .pdf, other: .tiff, .png, .xlsx, .eps, .py, .ipynb, .R, .Rmd | < 100 MB | |
| Project code and generated data – MACRO | Preprocessing, model code, evaluation code, model results | Generate new data | Digital | Software, other | .csv, .pdf, other: .tiff, .png, .xlsx, .eps, .py, .ipynb, .R, .Rmd | < 100 MB | |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| Project code and generated data - COLUMBUS | Preprocessing, model code, evaluation code, model results | Generate new data | Digital | Software, other | .csv, .pdf, other: .tiff, .png, .xlsx, .eps, .py, .ipynb, .R, .Rmd | < 100 MB | |
| Project code and generated data - BACE | Preprocessing, model code, evaluation code, model results | Generate new data | Digital | Software, other | .csv, .pdf, other: .tiff, .png, .xlsx, .eps, .py, .ipynb, .R, .Rmd | < 100 MB | |

*GUIDANCE:*

*DATA CAN BE DIGITAL OR PHYSICAL (FOR EXAMPLE BIOBANK, BIOLOGICAL SAMPLES, …). DATA TYPE: DATA ARE OFTEN GROUPED BY TYPE (OBSERVATIONAL, EXPERIMENTAL ETC.), FORMAT AND/OR COLLECTION/GENERATION METHOD.*

*EXAMPLES OF DATA TYPES: OBSERVATIONAL (E.G. SURVEY RESULTS, SENSOR READINGS, SENSORY OBSERVATIONS); EXPERIMENTAL (E.G. MICROSCOPY, SPECTROSCOPY, CHROMATOGRAMS, GENE SEQUENCES); COMPILED/AGGREGATED DATA[5] (E.G. TEXT & DATA MINING, DERIVED VARIABLES, 3D MODELLING); SIMULATION DATA (E.G. CLIMATE MODELS); SOFTWARE, ETC.*

*EXAMPLES OF DATA FORMATS: TABULAR DATA (.POR,. SPSS, STRUCTURED TEXT OR MARK-UP FILE XML, .TAB, .CSV), TEXTUAL DATA (.RTF, .XML, .TXT), GEOSPATIAL DATA (.DWG,. GML, ..), IMAGE DATA, AUDIO DATA, VIDEO DATA, DOCUMENTATION & COMPUTATIONAL SCRIPT.*

*DIGITAL DATA VOLUME: PLEASE ESTIMATE THE UPPER LIMIT OF THE VOLUME OF THE DATA PER DATASET OR DATA TYPE.*

*PHYSICAL VOLUME: PLEASE ESTIMATE THE PHYSICAL VOLUME OF THE RESEARCH MATERIALS (FOR EXAMPLE THE NUMBER OF RELEVANT BIOLOGICAL SAMPLES THAT NEED TO BE STORED AND PRESERVED DURING THE PROJECT AND/OR AFTER).*

[5] These data are generated by combining multiple existing datasets.

| | |
|---|---|
| If you reuse existing data, please specify the source, preferably by using a persistent identifier (e.g. DOI, Handle, URL etc.) per dataset or data type. | 1) IMPACT: dataset gathered in the IMPACT trial (ClinicalTrials.gov identifier: NCT02164513), not public, received under data sharing agreement via ClinicalStudyDataRequest.com.<br>2) PIONEER: dataset gathered in the PIONEER trial (ClinicalTrials.gov identifier: NCT02986321), not public, received under data sharing agreement.<br>3) ISAACC: dataset gathered in the ISAACC trial (ClinicalTrials.gov identifier: NCT01335087), not public, will be received under data sharing agreement.<br>4) BeNeBio: dataset gathered in the BeNeBio trial (ClinicalTrials.gov identifier: NCT04340076), not public, will be received under data sharing agreement.<br>5) MACRO: dataset gathered in the MACRO trial (ClinicalTrials.gov identifier: NCT00325897), not public, received under data sharing agreement.<br>6) COLUMBUS: dataset gathered in the COLUMBUS trial (ClinicalTrials.gov identifier: NCT00985244), not public, received under data sharing agreement.<br>7) BACE: dataset gathered in the BACE trial (ClinicalTrials.gov identifier: NCT02135354), not public, trial performed by our own lab. |

| | |
|---|---|
| Are there any ethical issues concerning the creation and/or use of the data (e.g. experiments on humans or animals, dual use)? If so, please describe these issues further and refer to specific datasets or data types when appropriate. | ☒ Yes, human subject data<br>☐ Yes, animal data<br>☐ Yes, dual use<br>☐ No<br>If yes, please describe:<br><br>The IMPACT, PIONEER, ISAACC, BeNeBio, MACRO, COLUMBUS and BACE trial datasets contain personal data from patients that was gathered in clinical trials. The data has been collected with informed consent and in line with the clinical study protocols. This project will use artificial intelligence for secondary analyses of these datasets.<br><br>Three of these datasets were (partially) collected outside of the EU:<br>1) IMPACT was a global multicentre study, with data recorded in approximately 1200 study centres in 37 different countries. Countries included are the USA, Argentina, Australia, Austria, Belgium, Brazil, Canada, Chile, China, Colombia, Czechia, Denmark, Finland, France, Germany, Hong Kong, Israel, Japan, the Republic of Korea, the Netherlands, New Zealand, Norway, Peru, Philippines, Poland, Puerto Rico, Romania, the Russian Federation, Singapore, South Africa, Spain, Sweden, Thailand, Turkey, Ukraine, the United Kingdom and Vietnam.<br>2) MACRO data was recorded in 17 sites associated with 12 academic health centres in the USA.<br>3) PIONEER data was recorded in Bulgaria, Germany, Hungary, Poland, Russia, Ukraine and the United Kingdom. |

| | |
|---|---|
| Will you process personal data[6]? If so, briefly describe the kind of personal data you will use. Please refer to specific datasets or data types when appropriate. If available, add the reference to your file in your host institution's privacy register. | ☒ Yes<br>☐ No<br>If yes:<br><br>- Short description of the kind of personal data that will be used:<br><br>The IMPACT, PIONEER, ISAACC, BeNeBio, MACRO, COLUMBUS and BACE trial datasets, that we will use for secondary analyses, contain personal data from patients that was gathered in clinical trials. Datasets contain patient characteristics (txt), lab values (txt) and results from clinical tests (txt). Datasets are fully anonymized before transfer to our research group.<br><br>- Privacy Registry Reference:<br><br>1) IMPACT: Ethical approval for secondary analysis of the datasets was requested via KU Leuven's PRET platform (PRivacy and EThics). Ethical approval was granted:<br>Privacy Registry Reference G-2022-5402<br>2) MACRO + COLUMBUS: Ethical approval for secondary analysis of the datasets was requested via KU Leuven's PRET platform. Ethical approval was granted:<br>Privacy Registry Reference G-2022-4831-R2(AMD)<br><br>3) ISAACC: Contracts for access to this dataset are in progress, ethical approval will be requested via KU Leuven's PRET platform.<br>4) BeNeBio: This trial will finish in januari 2024, after which contracts for access to the data will be drawn up and ethical approval will be requested via KU Leuven's PRET platform.<br><br>5) BACE: Trial from our own research lab, approved by UZ Leuven ethics committee (s55829) and the competent authority (Federaal Agentschap voor Geneesmiddelen en Gezondheidsproducten). |

[6] See Glossary Flemish Standard Data Management Plan

| | |
|---|---|
| | 6) PIONEER: details not fully know. |
| Does your work have potential for commercial valorization (e.g. tech transfer, for example spin-offs, commercial exploitation, …)?<br>If so, please comment per dataset or data type where appropriate. | ☐ Yes<br>☒ No<br>If yes, please comment: |
| Do existing 3rd party agreements restrict exploitation or dissemination of the data you (re)use (e.g. Material/Data transfer agreements, research collaboration agreements)?<br>If so, please explain to what data they relate and what restrictions are in place. | ☐ Yes<br>☒ No<br>If yes, please explain: |
| Are there any other legal issues, such as intellectual property rights and ownership, to be managed related to the data you (re)use?<br>If so, please explain to what data they relate and which restrictions will be asserted. | ☐ Yes<br>☒ No<br>If yes, please explain: |

## 3. Documentation and Metadata

| | |
|---|---|
| Clearly describe what approach will be followed to capture the accompanying information necessary to keep **data understandable and usable**, for yourself and others, now and in the future (e.g. in terms of documentation levels and types required, procedures used, Electronic Lab Notebooks, README.txt files, Codebook.tsv etc. where this information is recorded). | We receive the trial datasets (IMPACT, PIONEER, ISAACC, BeNeBio, MACRO, COLUMBUS and BACE) under data sharing agreements, together with all necessary dataset documentation.<br><br>The developed code will be well documented using general README.txt files, inline comments and function/class input – output explanations. |
| Will a metadata standard be used to make it easier to **find and reuse the data**?<br><br>If so, please specify which metadata standard will be used. If not, please specify which metadata will be created to make the data easier to find and reuse.<br><br>*Repositories could ask to deliver metadata in a certain format, with specified ontologies and vocabularies, i.e. standard lists with unique identifiers.* | ☒ Yes<br>☒ No<br>If yes, please specify (where appropriate per dataset or data type) which metadata standard will be used:<br><br>The IMPACT and PIONEER datasets follow the ADaM standards for analysis datasets and associated metadata.<br><br>If no, please specify (where appropriate per dataset or data type) which metadata will be created:<br><br>Dataset documentation for the other trial datasets (ISAACC, BeNeBio, MACRO, COLUMBUS and BACE) is available but not standardized. The developed code will be well documented using general README.txt files, inline comments and function/class input – output explanations. |

| 4. Data Storage & Back-up during the Research Project | |
|---|---|
| Where will the data be stored? | Trial datasets (IMPACT, PIONEER, ISAACC, BeNeBio, MACRO, COLUMBUS and BACE) and generated data are stored in a personal map on the KU Leuven J-drive (Shared network storage).<br>Code files are stored in private repositories on the KU Leuven GitLab server. |

| | |
|---|---|
| How will the data be backed up?<br><br>*What storage and backup procedures will be in place to prevent data loss? Describe the locations, storage media and procedures that will be used for storing and backing up digital and non-digital data during research.*[7]<br><br>*Refer to institution-specific policies regarding backup procedures when appropriate.* | An automatic back-up of the shared network storage is made according to the following schedule:<br>- A backup at 8:00, 12:00, 16:00 and 20:00, the last 6 of which are stored on ICTS servers in a KU Leuven datacentre.<br>- A daily backup, at midnight, the last 6 of which are stored on ICTS servers in a KU Leuven datacentre.<br>- A weekly backup, Saturday night at midnight, the last 12 of which are stored on ICTS servers in a KU Leuven datacentre.<br>Every hour, a mirror (exact copy) of all data is created in a second ICTS data centre. In the event that the primary storage unit is corrupted, the ICTS team can get the mirror copy online within the hour.<br><br>The KU Leuven GitLab server is a self-hosted platform: all data is stored in datacentres of KU Leuven with regular back-ups. |
| Is there currently sufficient storage & backup capacity during the project? If yes, specify concisely. If no or insufficient storage or backup capacities are available, then explain how this will be taken care of. | ☒ Yes<br>☐ No<br>If yes, please specify concisely:<br><br>There is no hard limit to the storage capacity of the KU Leuven shared network storage. A storage request can be made to ICTS, additional storage can be requested if needed.<br><br>Standard storage capacity for KU Leuven personnel on the KU Leuven GitLab server consists of 25 personal project repositories and 1GB storage for personal project repositories. We do not exceed this standard capacity.<br><br>If no, please specify: |

[7] Source: Ghent University Generic DMP Evaluation Rubric: https://osf.io/2z5g3/

| | |
|---|---|
| How will you ensure that the data are securely stored and not accessed or modified by unauthorized persons?<br><br>*Clearly describe the measures (in terms of physical security, network security, and security of computer systems and files) that will be taken to ensure that stored and transferred data are safe.* [7] | Our shared network storage is only accessible by our own research group, access is secured by the central KU Leuven login with Multi-Factor Authentication. Access to individual trial datasets (IMPACT, PIONEER, ISAACC, BeNeBio, MACRO, COLUMBUS and BACE) is secured by passwords with AES-256 encryption.<br><br>Code is stored in private repositories on the KU Leuven GitLab server, access is secured by the central KU Leuven login with Multi-Factor Authentication. |
| What are the expected costs for data storage and backup during the research project? How will these costs be covered? | The shared network storage costs € 503.66 per TB, per year. For this project we need less than 3GB. The cost per year for 3GB would be € 1.51.<br><br>Standard storage capacity for KU Leuven personnel on the KU Leuven GitLab server (which we do not exceed) is free of charge. |

| **5. Data Preservation after the end of the Research Project** | |
|---|---|
| Which data will be retained for at least five years (or longer, in agreement with other retention policies that are applicable) after the end of the project? In case some data cannot be preserved, clearly state the reasons for this (e.g. legal or contractual restrictions, storage/budget issues, institutional policies...). | Code, generated data and the COLUMBUS, MACRO and BACE trial datasets will be retained for at least five years.<br>The dataset from the IMPACT trial will be deleted after a two-year period as agreed with the data provider (data sharing agreement with GSK (GlaxoSmithKline).<br>For the ISAACC and BeNeBio trial datasets the period of retention will depend on the data sharing agreements (agreements still in progress).<br>PIONEER data will be not stored after the end of this year. |
| Where will these data be archived (stored and curated for the long-term)? | Trial datasets (IMPACT, PIONEER, ISAACC, BeNeBio, MACRO, COLUMBUS and BACE) and generated data will be archived on the KU Leuven J-drive (Shared network storage). Code will be archived on the KU Leuven GitLab server. |

| | |
|---|---|
| What are the expected costs for data preservation during the expected retention period? How will these costs be covered? | Costs are the same as for data storage and back-up during the project. |

| 6. Data Sharing and Reuse | |
|---|---|
| Will the data (or part of the data) be made available for reuse after/during the project? Please explain per dataset or data type which data will be made available.<br><br>*NOTE THAT 'AVAILABLE' DOES NOT NECESSARILY MEAN THAT THE DATA SET BECOMES OPENLY AVAILABLE, CONDITIONS FOR ACCESS AND USE MAY APPLY. AVAILABILITY IN THIS QUESTION THUS ENTAILS BOTH OPEN & RESTRICTED ACCESS. FOR MORE INFORMATION: HTTPS://WIKI.SURFNET.NL/DISPLAY/STANDARDS/INFO-EU-REPO/#INFOEUREPO-ACCESSRIGHTS* | ☐ Yes, in an Open Access repository<br>☒ Yes, in a restricted access repository (after approval, institutional access only, …)<br>☐ No (closed access)<br>☐ Other, please specify: |
| If access is restricted, please specify who will be able to access the data and under what conditions. | We are reusing trial datasets (IMPACT, PIONEER, ISAACC, BeNeBio, MACRO, COLUMBUS and BACE) retrieved from restricted access repositories. We can access the data under the conditions in a data sharing agreement for each dataset. Data sharing policy is decided by each data provider. |
| Are there any factors that restrict or prevent the sharing of (some of) the data (e.g. as defined in an agreement with a 3rd party, legal restrictions)? Please explain per dataset or data type where appropriate. | ☒ Yes, privacy aspects<br>☒ Yes, intellectual property rights<br>☒ Yes, ethical aspects<br>☐ Yes, aspects of dual use<br>☐ Yes, other<br>☐ No<br><br>If yes, please specify:<br><br>Restrictions based on granted ethical approvals and data sharing agreements for all trial datasets (IMPACT, PIONEER, ISAACC, BeNeBio, MACRO, COLUMBUS and BACE). |

| | |
|---|---|
| Where will the data be made available?<br>If already known, please provide a repository per dataset or data type. | NA |
| When will the data be made available?<br><br>*This could be a specific date (dd/mm/yyyy) or an indication such as 'upon publication of research results'.* | NA |
| Which data usage licenses are you going to provide? If none, please explain why.<br><br>*A data usage license indicates whether the data can be reused or not and under what conditions. If no licence is granted, the data are in a grey zone and cannot be legally reused. Do note that you may only release data under a licence chosen by yourself if it does not already fall under another licence that might prohibit that.*<br><br>*Example Answer: E.g. "Data from the project that can be shared will be made available under a Creative Commons Attribution license (CC-BY 4.0), so that users have to give credit to the original data creators."* [8] | NA |
| Do you intend to add a PID/DOI/accession number to your dataset(s)? If already available, please provide it here.<br><br>*Indicate whether you intend to add a persistent and unique identifier in order to identify and retrieve the data.* | ☐ Yes<br>☒ No<br>If yes: |

[8] Source: Ghent University Generic DMP Evaluation Rubric: https://osf.io/2z5g3/

| | |
|---|---|
| What are the expected costs for data sharing? How will these costs be covered? | NA |

| 7. Responsibilities | |
|---|---|
| Who will manage data documentation and metadata during the research project? | **PI Maarten De Vos** |
| Who will manage data storage and backup during the research project? | **PI Maarten De Vos** |
| Who will manage data preservation and sharing? | **PI Maarten De Vos** |
| Who will update and implement this DMP? | **PI Maarten De Vos** |