Plan Overview

A Data Management Plan created using DMPonline

Title: Onderzoek BMW - leerlijn Data Science - update en voorbeeld

Creator: John Meeuwsen

Principal Investigator: Rens van de Schoot, Marc van Mil, Annelies Pieterman-Bos

Data Manager: John Meeuwsen

Affiliation: UMC Utrecht

Template: UMC Utrecht DMP with DPIA V.3.0

ORCID iD: 0000-0001-7736-2091

ORCID iD: 0000-0002-7608-0014

ORCID iD: 0000-0003-2096-0044

Project abstract:

The goal of this study is twofold. First, we aim to describe the similarities and differences between the intended and implemented curriculum of our Data Science learning progression. Second, we want to evaluate the similarities and differences between the intended and perceived curriculum. Together they will help to describe characteristics of a learning progression that have the potential to increase data science skill retention and transfer. These insights can be used 1) for further development of our curriculum, 2) by other teachers and policy advisors to implement Data Science education in other degrees of Biomedical Sciences and related degrees, and 3) to provide theoretical insights into learning of Data Science knowledge and skills.

ID: 175980

Start date: 10-04-2023

End date: 17-05-2025

Last modified: 12-06-2025

Onderzoek BMW - leerlijn Data Science - update en voorbeeld

1. General features

1.1. Acronym/short study title

BMWB-Data_Science_Leerlijn

1.2 Division of Principal Investigator

- Onderwijscentrum (Education Center)

Education center, Bachelor Biomedical Sciences

1.3 Department

Education center, Bachelor Biomedical Sciences

1.4 Path of the Research Folder

\\ds\data\DOO\Onderzoek

1.5 WMO/DEC

- non-WMO

Non-WMO research is research that is not part of the law on medical scientific research on humans (Wet medisch wetenschappelijk onderzoek met mensen (WMO)).
Most educational research is not medically oriented.

1.6 Research type(s)

- Qualitative
- Use of questionnaires

For this research we use:

- data and notes from the development of the learning trajectory;
- educational data (exam results);
- questionnaires completed by students;

- focus groups with students;
- focus groups with teachers.

1.7 Research design(s)

- Retrospective
- Observational

We performed the research at the end of the learning trajectory, where students finished the last obligatory course. In the questionnaire we ask them how they have experienced Data Science education.

1.8 Mono or multicenter study (one choice)

- Monocenter

1.10 Which organization is the sponsor of the study?

Sponsor is not applicable.

1.11 Name of datamanager consulted

John Meeuwsen

1.12 Last check date by datamanager

2025-04-16

1.13 Indicate which laws and regulations are applicable for the project (please check all that apply)

- Nederlandse gedragscode wetenschappelijke integriteit
- Richtlijn Kwaliteitsborging Mensgebonden Onderzoek (Quality Assurance for Research Involving Human Subjects)
- Algemene Verordening Gegevensbescherming (AVG) or General Data Protection Regulation (GDPR)

For most educational research, the following laws and regulations are important:

- GDPR, due to the use of privacy sensitive information.
- Quality Assurance for Research Involving Human Subjects

- Nederlandse gedragscode wetenschappelijke integriteit

2. Data Collection

2.1 Give a short description of the research data.

Subjects	Volume	Data Source	Data Capture Tool	File Type	Format	Storage space
Human	100	Questionnaires	Qualtrics, Excel and Rstudio	quantitative and qualitative	.csv and .docx	0-1 GB
Human	30	Focus group*	Recorded on dictaphone macbook**, transcribed and analyzed in Atlas and word	qualitative	.mp3 and .docx	0-10 GB
Human	50	TestVision	Excel and Rstudio	quantitative	.xlsx and .r	0-1 GB

* We wanted to perform five or six focus groups with five or six students, but in practice, it appeared to be difficult to organise meetings where enough students were available. Therefore, the focus groups were organised with groups of two students.

** Note that the recording on a device can be transferred to the cloud (iCloud or OneDrive) and that it is recommended to stop synchronising during the recording. After recording, files are moved to the Research Folder Structure. After movement, the local copies are deleted.

2.2 Describe the flow of the data (name systems used and/or third parties, recipients) <add link to location where diagram is stored in RFS>

- Questionnaires are completed in Qualtrics
- Answers to questionnaires are exported to Excel (csv)
  - Qualitative answers are analyzed in Atlas
  - Quantitative answers are analyzed in R
- Testvision results are reported and/or analyzed in Excel/R
- Focus groups are recorded and transcribed
  - Analyzed in Atlas

2.3 Estimated storage space for your project

- < 250 GB (e.g. questionnaires, textfiles, datasets)

2.4 Can you reuse existing data? If so, list the data source(s)

- Other, describe below
- No, please specify below

Partly yes, partly no:

Yes:
We use pseudonymized data that is previously collected in the course 'Bio-informatica' of the Bachelor Biomedical Sciences. Besides, we use notes and course materials to evaluate the intended and the implemented curriculum.

No:
The other data is not available. The DS learning trajectory in the BSc Biomedical Sciences is newly developed, and the findings are dependent on the context. Therefore new data using questionnaires and focus groups has to be collected in order to comprehensively evaluate the learning trajectory.

2.5 Describe how you will take care of good data quality.

Data from the test will be collected using TestVision. Questionnaires will be performed with Qualtrics, skips and validation chekcs are built in. Data quality will be checked by the researchers. Incomplete questionnaires will not be used.

During analysis of the focus groups, at least a part of the analyses will be performed by two researchers. In addition, researchers keep a logbook with their perceptions of research (as it may change over time).

#	Question	Yes	No	N/A
1.	Do you use a GCP-compliant Data Capture Tool or Electronic Lab Notebook?	X		
2.	Have you built in skips and validation checks?	X		
3.	Do you perform repeated measurements?			X
4.	Are your devices calibrated?			X
5.	Are your data (partially) checked by others (4 eyes principle)?	X		
6.	Are your data fully up to date?	X		
7.	Do you lock your raw data (frozen dataset)	X		
8.	Do you keep a logging (audit trail) of all changes?	X		
9.	Do you have a policy for handling missing data?	X		
10.	Do you have a policy for handling outliers?	X		

Short explanation for each step:

- (1) Data capture tool
  - Qualtrics (based on collaboration between UU and UMC) is GDPR compliant.
  - Recorded focus groups are recorded on MacBook - dictafoon. After recording the files are transferred to the Research Folder Structure (RFS) and the local copies were deleted.
- (2) Skips and validation checks
  - Cronbach alpha values for different questions from the same factor (for example five questions about internal motivation) were calculated to validate the used factors.

- (3) Repeated measurements
  - We only measured the experience of students once. Not applicable here, as we don't compare measures within students.
- (4) Calibration devices
  - In our research, the calibration of devices is not applicable.
- (5) Data (partially) checked by others
  - Yes, the two researchers checked parts of each others data. For example, the exclusion of samples, the scripts and analyses, and the thematic analyses were checked.
- (6) Data up to date
  - Yes, data was very recent, newest available data was used.
- (7) Raw data (frozen dataset)
  - After collection of data (recording, questionnaire), a dataset without any manipulation was stored in the RFS.
- (8) Logging (audit trail)
  - A log book was used to describe the changes and analyses for the thematic analyses. Also, for quantitative analyses, a log book was used and analyses were described in an RMarkdown file.
  - The level of detail can be hard to determine for such a log book.
- (9) Policy for handling missing data
  - If questionnaires are not completed, we still use the data as much as possible. Only if different questions, belonging to one factor, are not available, these data will not be used.
- (10) Policy for handling outliers
  - Outliers will be included in the analyses, unless there is a proper reason to exclude outliers. This will be determined by researchers Annelies and John with noting a clear reason.

2.6 Specify data management costs and how you plan to cover these costs.

#	**Type of costs**	**Division** ("overhead")	**Department**	**Funder**	**Other** (specify)
1.	Time of datamanager	X			
2.	Data capture tool	X			
3.	Storage	X			
4.	Archive	X			
5.					

2.7 Please give some more details on other centers and organizations involved. What are the roles of the other centers and organizations involved? (What research activity does this organization carry out in relation to the study and the data?)

There are no other centers and organisations involved.

2.8 Which contracts are in place?

There are no contracts in place.

2.9 State how ownership of the data and intellectual property rights (IPR) to the data will be managed

UMC Utrecht and UU are the owners of all collected data for this study. It's a collaboration between UMCU and UU as researchers are employed in both places.
The data is collected within a student group. Our data cannot be protected with IPR, but its value will be taken into account when making our data available to others, when setting up Research Collaborations and when drawing up Data Transfer Agreement(s).

2.10 Use of new technology. Does your study involve the implementation of a technology that has not been used before at UMC Utrecht?

- No

2.12 Will the study need/use personal data (directly or indirectly identifying)? For example, from the Electronic Patient Files (EPD; HiX), DNA, body material, images or any other form of personal data?"

- Yes. You have indicated that you are using personal data in your project. The following chapter is the Data Protection Impact Assessment (DPIA) for research data. It is derived from the full DPIA, in accordance with the privacy office of UMC Utrecht. Answering questions in this chapter helps to determine the risk of processing the personal data and what measures to take to minimize these risks.

We have information about grades from students and their gender.

3. Data Protection Impact Assessment (DPIA)

3.1 Describe the recipients outside the UMC Utrecht to whom the personal data are provided, what their role is (controller or processor) and where they are located.

- All systems and service providers involved are mentioned in question 2.1 and 2.2. All of them are already contracted by UMC Utrecht. I do not share personal data with other organisations.

3.4 What type of sensitive personal data will be used?

None of the above mentioned data will be used.
Data about study progress and study performance is used.

3.5 What type of directly or indirectly identifying personal data will be used? Indicate why you need this data. Is this truly necessary?

Remove the data points you are not processing. Examples are here to guide you, make sure to specify the exact data point. Add the data points that are not mentioned here yet.

Category of personal data	**Reason for collecting these data**
Research parameters	Research protocol chapter(s): These datapoints answer my specific research question. The research question cannot be answered without these data points. The research datapoints do/do not involve sensitive data.
Name	Name is used to contact students and to plan interviews with them.
Address	NA
Telephone number	NA
Email	E-mail is used to contact students and to plan interviews with them.
Age (not categorized)	NA
Date of birth	NA
Gender	Yes, to identify possible differences in study performance between male and female students. Without this data the Research Question can't be answered.
Imaging e.g. MRI, pictures or video (can be health data)	NA
Sound recordings (may be health data)	Interviews are recorded with the dictaphone of the Macbook. To prevent uploading data to the cloud, synchronisation of the data is paused.
Location data (e.g. postal code)	NA
Personal interests	NA
Financial data (e.g. bank account number)	NA
Other datapoints that are not yet mentioned: Datapoint 1	Student number and study achievement: Without those, the research question can't be answered.
Datapoint 2	What purpose will be achieved with this datapoint?:
Datapoint 3	What purpose will be achieved with this datapoint?:

3.6 Select any vulnerable groups from which you will collect data.

- Other --> describe
- Employees

Students and teachers are both interviewed to collect data about the experienced learning trajectory.

3.7 Which legally prescribed personal number will be used? Note: it is NOT allowed to use BSN (or its international counterpart) for scientific research purposes.

- None

3.8 Can the purpose of the study be achieved with anonymous or pseudonymized data?

- No, I need direct identifying personal data to answer the Study research question the dataset is stored in folder C_PersonalData of your research folder structure with access only for the persons that need access to this data (explain why you cannot do the research without this data)

The student number is needed to couple the questions of the questionnaire to the grades for the courses.

3.9 Which measures are taken to prevent the data from being traceable to the natural person? Also consider the measures taken to prevent data breaches.

- SOPs about who and how an employee has access
- 2FA/MFA before access to (health) data
- Aggregation of data
- Clear retention period(s)
- Role specific access to identifying data
- Minimalization of collected data points
- Pseudonymization of data

3.10 Does the reuse of the data fit within the purpose for which they were originally collected?

- Not applicable, we will not reuse data

3.11 Are data subjects contacted and included only after informed consent?

- Yes, we ask study-specific or other type of Informed consent (e.g. broad consent, deferred consent).

All data subjects are included (for the questionnaire and interview) after they have given informed consent.
For the quantitative data, only aggregated data was used for all students.

3.16 What type of consent for using personal data is obtained?

- Study-specific or other type of Informed consent (e.g. broad consent, deferred consent, explain).

Specific consent for this study is asked.

3.17 Is there a dispute settlement or a party where the subject can go to with questions or complaints about the processing of personal data?

- Subjects are provided contact information whom and how to contact the study team via the PIF. Also, subjects are informed about their possibility to contact the data protection officer (DPO) or supervisory authority (Autoriteit Persoonsgegevens).

In the information letter, contact details of the data protection officer are provided.

3.18 Describe how you manage your data to comply to the rights of study participants.

- A subject can object to processing of their personal data or withdraw consent
- We inform the subjects about their rights of access, rectification and deletion of their data. In the information provision we describe the contact information in case a subject wants to exercise their rights,

3.19 Does the data collected concern data from which behavior, presence or performance (profiling) can be measured when this is not the purpose of the research?

- No

3.20 Are automated (i.e. without any human intervention) decisions made about the subjects based on the data?

- No

3.21 Describe the tools, procedures and transport methods that you use to ensure that only authorized people have access to personal data

- We use the secured Research Folder Structure that ensures that only authorized personnel has access to personal data, including the key table that links personal data to the pseudoID

3.22 Describe your backup strategy or the automated backup strategy of your storage locations.

- All (research) data is stored in the RFS on UMC Utrecht networked drives from which backups are made automatically twice a day by the division IT (dIT).

3.23 Describe who will have access to which data during your study.

Type of data	Who has access
Grades	Main researcher who is also teacher, Data manager
Interview data and open questions of questionnaire	Main researcher, data manager
Pseudonomized data	Research team

3.24 Indicate the ISO who was consulted for this DPIA and what advice follows from this?

5. Metadata and Documentation

5.1 Describe the metadata that you will collect and which standards you use.

We do not use metadata standards yet. The data will be delivered including a data dictionary. For every variable this data dictionary contains an explanation of the values.
Other metadata we collect and describe (either in methods section of publication or as document in the RFS and datapackage):

- Description of participants (year of study, age range, etc.)
- Description of educational context (program, university, level, expected career paths, etc.)
- Timing of data collection (in program, in relation to exams, etc.)
- Researcher positionality (background, relation to student, etc.)
- Oorsprong van vragen uit vragenlijst (oorspronkelijke vragenlijst, Engelse vragen)
- Qualitative analysis codebook

5.2 Describe your version control and file naming standards.

We distinguish versions by indicating the version in the filename of the master copy by adding a date and if necessary a code after each edit, for example 2023-11-22_filedescription_V1.1 (first number for

major versions, last for minor versions). The most recent copy at the master location is always used as the source, and before any editing, this file is saved with the new version code in the filename. The file with the highest code number is the most recent version and older versions are moved to a folder OLD.

6. Data Analysis

6 Describe how you will make the data analysis procedure insightful for peers.

- It is anticipated that we are going to write a paper and publish it, which will make the research accessible to peers.
- For each experiment, I will document my analysis steps in an Electronic Lab Notebook (ELN). My documentation is shared with other members in the same research group. In future publications, relevant analysis steps will be described and available for everyone.
- I will make an overview of datasets and analysis scripts, such that it is fully clear how the statistical analysis is performed. Peers will be able to repeat the analysis based on my overview.
- We will be using tools like SAS, R or SPSS for statistical analysis of the data. The scripts will contain comments, such that every step in the analysis is documented and peers can read why I made certain decisions during the analysis phase.
- I have written an analysis plan in which I state why I will use which data and which statistical analysis we plan to do in which software. The analysis plan is stored in the project folder, so it is findable for my peers.

The analysis plan, preregistration R scripts, final analysis scripts (RMarkdown), qualitative analysis codebook, pre-analysis and post-analysis educational design conjecture map (analytical output), questionnaire, interview guide and data dictionary are available on the Open Science Framework: https://osf.io/rwnb8/files/osfstorage, DOI 10.17605/OSF.IO/RWNB8.

The researcher logbook, detailing steps of the analysis, is stored in the Research Folder Structure, and used to describe relevant analysis steps in the publication.

7. Data Preservation and Archiving

7.1 Describe which data and documents are needed to reproduce your findings.

The data package will contain:

- the raw data
- the study protocol describing the methods and materials
- ethical review board application
- ethical review board approval
- researcher logbook with audit trail
- the questionnaire used with sources for each question
- the script to process the quantitative data
- the scripts leading to tables and figures in the publication
- a data dictionary with explanations on the variable names

- transcription procedure manual
- the final codebook used for qualitative analysis
- ‘read_me.txt’ file with an overview of files included and their content and use.

7.2 Describe which archive or repository (include the link!) you will use for long-term archiving of your data and whether the repository is certified.

- After finishing the project, the data package will be stored at the UMC Utrecht Research Folder Structure and is under the responsibility of the Principal Investigator of the research group. The (meta)data will be published in DataverseNL, the preferred UMCU repository.

The full data package will be stored at the UMC Utrecht Research Folder Structure. Within that RFS, a copy will be made that can be shared outside of the project team. Our participants have been explicitly asked for consent to reuse their data for other studies with similar design and aims, in a separate question in the consent form. Therefore, a separate data file will be created that includes only data of participants that have given their consent for this reuse. The metadata of this study and datafile will be published in DataverseNL. Other researchers can request access to this filtered data file. They will be asked to sign a data transfer agreement to obtain the data file.

7.3 Give the Persistent Identifier (PID) that you will use as a permanent link to your published dataset.

- A PID will be generated when a data package is published on DataverseNL. This PID will be updated when available in the additional comment area of this plan.

The PID will refer to the metadata of the data package that is available after signing a data transfer agreement.

8. Data Sharing Statement

8.1 Describe what reuse of your research data you intend or foresee, and what audience will be interested in your data.

Given the personal characteristics of the data and the context-specificity, it will be hard to reuse the research data. Direct colleagues might want to reuse them to generate or answer new research questions related to data science learning in biomedical sciences.

8.2 Are there any reasons to make part of the data NOT publicly available or to restrict access to the data once made publicly available?

- Yes (please specify)

As the data is privacy-sensitive, we publish the descriptive metadata in the data repository with a description of how a data request can be made (by sending an email to one of the authors). In the event that peers like to reuse our data this can only be granted if the research question is in line with the original informed consent signed by the study participants. Every application therefore will be screened upon this requirement. If granted, a data usage agreement is signed by the receiving party.

8.3 Describe which metadata will be available with the data and what methods or software tools are needed to reuse the data.

Along with the publication, the following files will be made available:

- the focus group guides
- the questionnaire
- an explication of the sources and translations of questionnaire questions
- codebook of the qualitative data
- the scripts for the quantitative analysis in R (both preregistration and final script)
- a data dictionary

8.4 Describe when and for how long the (meta)data will be available for reuse

- (Meta)data will be available upon completion of the project

Data will be made available upon completion of the project and at least before publication of the article.

8.5 Describe where you will make your data findable and available to others.

Any (meta)data of the article will be made available through OSF:
https://doi.org/10.17605/OSF.IO/RWNB8
The preregistration can be found under the "Registrations" tab: hypotheses, study design, sampling, and methods.
(Meta)data can be found under the "Files" tab.