DMP

Please complete the survey for your data management plan below.

Thank you!

Response was added on 23-07-2020 15:55.

Project title*	Coronary ARtery disease: Risk estimations and Interventions for prevention and EaRly detection (CARRIER)
Grant number*	628.011.212

Project abstract*	CARRIER targets detection and primary and secondary prevention of coronary artery disease (CAD) with a regional alliance of clinicians, citizens, legal experts, and data scientists collaborating on research of big data-driven, participative self-care interventions. CAD is the most common cardiovascular disease and one of the leading causes of deaths and disability. Strong clinical evidence exists for the benefit of physical activity, healthy diet, and cessation of nicotine use. However, only a minority of citizens participate in rehabilitation programs to prevent CAD. Internet and smartphone-based self-care offers a wider reach for such interventions. CARRIER will combine clinical big data from different sources (hospitals and general practitioners) with socio- economic big data and artificial intelligence to build models that will drive detection and prevention of CAD with an intervention delivered via an electronic multimedia gamified lifestyle coach (eCoach). A prognostic model will help identify patients at increased risk (primary prevention) that along with patients with CAD (secondary prevention), will form the target population. The participants, together with clinicians, will co-create a personalised health management plan, and they will be supported by the eCoach to adhere to it. The use of the eCoach will generate data on the participants' lifestyle that will feed and validate a predictive model to estimate the personalised benefit of lifestyle changes. This will inform clinicians and will affect the behaviour of the eCoach. CARRIER will offer valuable insights into the effectiveness of eCoach-supported self-care for CAD and the value of clinical and socio-economic data in early detection of CAD.

1. PROJECT GROUP

1.1 Who is the contact person of the project? Please include full name, institute, e-mail address, telephone number and ORCID.*	ALAJ (Andre) Dekker Institute of Data Science, Maastricht University andre.dekker@maastro.nl +31 6 24102304 ORCID 0000-0002-0422-7996
1.2 Is there a person responsible for data management in this project?*	⊗ Yes ◯ No
1.3 Please provide full name, institute, e-mail adress, telephone number and ORCID of the data manager.*	Inigo Bermejo Clinical Data Science, Maastricht UMC+ inigo.bermejo@maastro.nl +32 474235784 ORCID 0000-0001-9105-8088

1.4 Is there a back-up data manager?*	⊗ Yes ◯ No
1.5 Please provide full name, institute, e-mail adress, telephone number and ORCID of the back-up data manager.*	Leonard Wee Clinical Data Science, Maastricht UMC+ leonard.wee@maastro.nl +31 6 875 38 145 ORCID 0000-0003-1612-9055
1.6 Who is the contact person after the project has ended? Please include full name, institute, e-mail adress, telephone number and ORCID.	ALAJ (Andre) Dekker Institute of Data Science, Maastricht University andre.dekker@maastro.nl +31 6 24102304 ORCID 0000-0002-0422-7996

2. DATA REUSE

2.1 Did you search for existing data that could be reused in your project? This could be either third-party data or data of your own institute.*	⊗ Yes ◯ No

Catalogues and repositories:

- BBMRI
- clinicaltrials.gov
- Dash
- Dataverse
- Figshare
- re3data
- UK data archive
- YODA
- Zenodo

2.3 Will you reuse existing data in the current project? This could be either third-party data or data of your own institute.*	⊗ Yes ◯ No

2.4 Did you check whether the informed consent form used for the existing data collection allows you to reuse this data? Describe how the intention of reuse is approached by the third party data, or simply copy and paste the statement from the informed consent form.*	We plan to use existing data from three data sources: CBS, general practitioners (HOZL, ZIO, MCC Omnes) and hospitals. CBS makes a number of their datasets available for research and policy making, through their microdata services. It is also possible to upload your own datasets and link these with CBS microdata. There is no mention on the necessity of informed consent. Regarding GP data, each patient representative organisation has their own rules regarding the reuse of data for research. However, informed consent is not usually required for research on anonymised data. Regarding hospital data, we are planning to work at least with data from our own institution, MUMC+, and maybe other hospitals (e.g. Zuyderland). The METC niet-WMO application process of MUMC+ describes the conditions in which informed consent is required for reuse of existing data. No permission is required for the further use of anonymised data (i.e. data that is not traceable to the person at all). The anonymous provision of patient data by the practitioner is therefore permitted; the researcher cannot trace the identity of those involved. In research where traceable data is further used, the starting point is that informed consent is requested. Test subjects must be informed about the research and must give permission for the use of their data. Exceptions are possible: -□ if it is not reasonably possible to request permission (eg if the patients have died or are unreachable) and safeguards have been taken that do not disproportionately harm the privacy of the patient; -□ if requesting permission cannot reasonably be required (eg if the investigation is very large or there is a significant risk of bias) and the information is provided in such a way that conversion cannot be reasonably prevented. In addition, the following conditions must be met: - the research has a general interest; - the research cannot be conducted without the relevant data; - the patient has not objected to the use of his / her data for scientific research;

2.5 What documents or websites exist describing the access, privacy, secondary use and co-authorship rules and policies for reusing the existing data?*	CBS: https://www.cbs.nl/en-gb/about-us/organisation/privacy https://www.cbs.nl/en-gb/our-services/customised-services- MUMC: https://www.mumc.nl/sites/default/files/research_code_broc GPs: HOZL: https://www.huisartsen-ozl.nl/privacy.html ZIO: https://www.zio.nl/wp-content/uploads/2018/06/infobrief-uw MCC Omnes: https://mcc-omnes.nl/privacy-policy
2.6 Who is the contact person responsible for the existing data? Please include full name, institute, e-mail adress, telephone number and ORCID.	CBS: Bob (R.D.) van den Berg rd.vandenberg@cbs.nl MUMC+: Pascal Suppers pascal.suppers@mumc.nl GPs: Not yet defined

3. CREATING AND PROCESSING DATA

3.1 What metadata will be produced on project-level? Will you use a metadata standard?*	We will use the Dublin Core Metadata element set for project level metadata.
3.2 Did you search for metadata standards at data level that could be used in the project? Please explain your answer. Include a description of the metadata standard, if applicable.*	We will search for ontologies and vocabularies to reuse concepts in order to avoid creating duplicates. The metadata standard that will be used will be RDF, based on our experience in past projects.
3.3 What kind of variables will be measured at the study? Specify which tools or instruments will be used for measuring the data. In which document will the variables be described in detail?*	Lifestyle variables: Diet, exercise, use of alcohol and tobacco, etc. These variables will be self reported through the eCoach. Clinical variables: Blood pressure, cholesterol, hospital attendance, cardiovascular events. These variables will be measured by nurses or GPs using medical instruments and stored in the hospital/GP EHR. The variables and the measurement process will be part of the knowledge graph describing the data. These variables will be described in detail in deliverable WP2.3 - Report describing data infrastructure framework (M9)
3.4 List all sources of data and include the following information (use the answer example as guidance, but feel free to include additional fields if relevant): o Data source; o Software necessary for reading it; o Version of the software; o Format of the data (e.g. .csv, .dat); o Where can the software be accessed. Example answer: Source: Tabular data with physiological and anthropometrical. Software: SPSS 22.0; R Statistics 4.2. Format: .sav; .csv. Software access: www.spss.com; www.r-project.org.	Data sources: - CBS - GP institutions - MUMC Format of the data and software necessary to read it is unknown at this point.

3.5 How are the datasets and raw data going to be named and stored? How will you name your folders? How will you structure your files? What naming conventions will you use? Include: o Name of the file (including date and version) o Description of the file o Where is the data stored Example answer: Description: Final dataset of processed data used for statistical analyses included in the publications Smith et al. 2019. Path and file name: G:/Department/StudyA/Analyses/20180629_EW01_FinalDataset.csv G:/Users/StudyA/Documents/Datasets/StudyA_FinalDataset_01012018.csv.	Data will be stored in a graph database using the GraphDB software, which handles versioning.
More information on file naming: See digitalscholarshipleiden.nl and GARP for an example.	
3.6 How will you secure a master file? How will you handle versioning?	
3.7 Will you create a data dictionary? Where is it going to be stored?	No data dictionary will be needed because all the variables will be mapped to a universal standard and publicly accessible lexicon such as SNOMED CT.
3.8 Will established terminologies or ontologies be used in the project? Which terminology/ontology will be used for which variable? Please explain your answer.	We will use a number of established vocabularies and terminologies. For clinical data we plan to use SNOMED CT, while for non-clinical data we will use different ontologies. For example, for sociodemographics data we will use CBS Linked Data ontology and we are looking for an appropriate ontology for eHealth applications. If we do not find appropriate ontology for our data, we will create one.
3.9 What will be the procedure to standardize variables without standard ontologies? Will a codelist be created? How will it be accessible for other researchers? Will it relate to established terminologies/ontologies?	N/A

4. DATA COLLECTION AND IT PROFESSIONALISM

4.1 How is existing data going to be combined with new data?*	We will combine data in two ways: data from different sources about the same individuals, and data about different individuals from different sources. For the former, we will generate identifiers based on common data elements so that we can match individuals across datasets. For the former, we will use semantic ontologies for data harmonisation, in order to make sure that we do not mix data referring to different concepts.

Data harmonization

Data harmonization can be useful for researchers using third party data, or for studies that will have data collected in multiple center.
Combining different datasets consists of pooling heterogeneous different data sets and transforming them into one merged and complete data set. There are many ways to conduct this procedure, such as making use of common variables (i.e. variables that are common to the different data sets such as age or sex) or by generating new variables from different items. These variables are entitled "common data elements" (Rolland et al. 2015).

The following articles describe approaches for data-harmonization:
Rolland B, Reid S, Stelling D, et al. Toward Rigorous Data Harmonization in Cancer Epidemiology Research: One Approach. American Journal of Epidemiology. 2015;182(12):1033-1038. doi:10.1093/aje/kwv133.

Fortier I, Burton PR, Robson PJ et al. Quality, quantity and harmony: the DataSHaPER approach to integrating data across bioclinical studies. Int J Epidemiol. 2010;395:1383-1393.

Fortier I, Doiron D, Little J et al. Is rigorous retrospective harmonization possible? Application of the DataSHaPER approach across 53 large studies. Int J Epidemiol. 2011;405:1314-1328.

Doiron D, Burton P, Marcon Y et al. Data harmonization and federated analysis of population-based studies: the BioSHaRE project. Emerg Themes Epidemiol. 2013;101:12.

4.2 How are data edits going to be documented?*	Edits on text files containing data will be documented using content versioning systems such as git. Edits on databases will be documented through the software (e.g. GraphDB), which allows the recovery of the content of the database at any point in time.

4.3 Is the data going to be audited/monitored?* Some possibilities are: - check completeness of records - perform in-depth checks for selected records - perform logical and consistency checks - automate checks whenever possible	⊗ Yes ◯ No

Data audits

Data audits can help improving the quality of the data. There are several guidelines for auditing data, such as the NCI Guidelines for Auditing Clinical Trials.

The Nederlandse Federatie van Universitair medisch centra (NFU) has also reported guidelines on data audits.
It should be indicated who is responsible for conducting the audits, and how audit report forms can be accesed.
These forms should be made acessible at a data repository at a further stage.

4.4 Who is responsible for conducting the audits and how can audit report forms be accessed?*	Inigo Bermejo and Leonard Wee will be respondible of conduction data audits. audit report forms can be accesed. Audit report forms will be made acessible in a data repository at a further stage.
4.6 Are there going to be strategies to prevent data entry mistakes? Please explain your answer.*	In the eCoach we will use validation rules that enforce a valid range of values and type of information is entered in each field.
4.7 How is the data going to be stored and backed-up during the data collection phase?*	DataHub Maastricht will manage the secure data storage and will apply the backup policies and schedules applicable in the case of sensitive clinical data.
4.8 What are the access rules of the database across the project members? (i.e. who owns read/write access, etc.)	_____
4.9 Is it needed to link multiple independently-collected data sets at the participant level?	⊗ Yes ◯ No
4.10 How is this data linkage performed in a privacy-sensitive manner? Is the data going to be anonymized or pseudonymized?	An identifier will be generated based on common variables and then encrypted so that it does not constitute a leak of personal information.

More information on anonymization and pseudonymization:

Anonymous data refers to data where re-identification is impossible.

Pseudonymous data is a form of de-identification, in which a part of personal information remains. This concept is not formally defined in the current EU data protection legal framework. (wsgrdataadvisor.com)

There are many techniques for data pseudonymization. The Working Party on the Protection of Individuals with Regard to the Processing of Personal Data has issued an opinion document on different anonymization and pseudonymization techniques. (ec.europa.eu)

Moreover, the UK Data Archive has issued guidelines on qualitative and quantitative data anonymization.

4.11 What is the procedure to harmonize different datasets? (e.g. third-party, multicenter)	_____
4.12 Which common data elements between the datasets are used for aiding data harmonization?	_____

4.13 Which (electronic) data capture software will be used for collecting the data (e.g. eCRF system or wearable device)? Does the software have any of the following specific features: o User logs (i.e. whether it registers user activity on the database, especially desirable for audits); o Data field validation (e.g. only allowing numbers or dates to be typed on a certain variable); o Using reference values (pre-defining minimum and/or maximum values allowed).	The eCoach will consist of a digital platform that will allow self reporting of lifestye activities. The eCoach will create user logs, will contain data field validation and reference values.
4.14 How are queries going to be generated and managed during the data cleaning process?	_____

5. PRIVACY AND INTEGRITY

5.1 Does the project need approval by a medical ethical committee, animal ethical committee, biobank committee or another ethical committee? To which committee was the approval requested? What is the current status? Please explain your answer.*	As part of the project, we will run at least two studies: one small usability study to fine-tune our intervention and a bigger one to assess the impact of the intervention on patients' health and on the healthcare system. In order to run these studies, we will seek approval from the METC of the MUMC+. The approval has not yet been requested as the details of both studies have yet to be defined.
5.2 Is it necessary for your project to obtain informed consent of the participants?*	⊗ Yes ◯ No
5.3 What is the procedure to obtain informed consent of the participants? Will you use paper or digital forms? How are the forms going to be made accessible at a later stage?*	Patients will fill in digital forms at recruitment points. The consent forms will be publicly available online and linked to the metadata generated in the project.
5.4 Does the informed consent state that data created in this project can be reused for new projects? Please explain your answer.*	Yes. In the form, we will seek consent from participants in the project for the reuse of pseudonimised data for research.
5.5 Is there a committee assigned to review privacy and integrity issues of the project? Please explain your answer.*	Yes, it will be an internal committee formed by clinicians, data scientists, data engineers and data privacy law experts, that will assess the privacy and integrity issues.
5.6 Will you use wearable devices?*	◯ Yes ⊗ No

Wearable devices

Data captured through wearable devices poses potential privacy and integrity risks related to its underlying processes. Besides, privacy concerns, the quality of the data might be threatened by attributability issues (e.g. the device being used by someone other than the participant) or technology related issues (such as data limit, lack of wireless connection, inadequate calibration). Many of these issues have been discussed in the paper eSource in Clinical Research: A Data Management Perspective on the Use of Mobile Health Technology, issued by the Society for Clinical Data Management.

5.8 How will sensitive data be handled to ensure it is stored and transferred securely? How will the identity of the participants be protected? Example answers: 1) Identifying data will remain within the institution and will never be published. 2) All data will be stored pseudonymized. 3) It is not possible to trace back patients. 4) All data will be sent through a secure connection.	Identifying data will remain within the institution and will never be published. Published data consist of aggregated data (i.e. summary statistics) so that it is not possible to trace back individual patients.

6. BUDGET

6.1 How will data management be costed in the project?*	Data management costs are included in the budget as "project-related goods/services". As some of the costs related to data have been accounted elsewhere, we will cost only the resources that would be needed to preserve and make research data shareable beyond our consortium. These will include the fees that the universoty data management department (Datahub) charges, that cover the human effort, infrastructure and tools needed to manage, store and provide access to data.

Costing data management

To cost research data management in advance can substantially reduce the costs of the project. The UK Data Archive provides a data management costing tool that can be helpful for answering this item.

6.2 Estimate the costs for sustainability (long-term storage). Make an estimation of the disk space needed for long-term storage of the data (after the data cleaning process). Indicate which (meta)data will be stored in long-term and where.	_____
6.3 Did you realize the cost estimation for data management and sustainability?	_____

7. DATA SHARING

7.1 Will you make data available for reuse at the end of your project?	⊗ Yes ◯ No
7.2 Which data is going to be made available for sharing and how is it going to be accessible (totally open, under co-authorship agreements, embargo period etc)? Access rules may vary between different types of data.	Given the limitations raised by GDPR, we will only make public aggregated data generated through the use of the eCoach. We would like to make this data completely open.
7.3 How are the datasets going to be made findable? Which catalogues and/or repositories will be used to place the (meta)data of the project?	Following the FAIR principles, we will assign a globally unique and persistent identifier to the dataset, we will describe the dataset with rich metadata and register it in a searchable resource such as the BBMRI repository.

7.4 How is your data going to be identified? List all the persistent identifiers used for data and documentation (e.g. DOI, ISBN).	_____
7.5 Are your metadata available and sufficient for other researchers to understand your data?	Our plan is that the metadata will describe the data with sufficient detail for independent researchers to understand it.
7.7 Which kind of formal documents (e.g. agreements, contracts) were produced? Where is it located and how can it be accessed? Which of these formal documents need to be retained /preserved for contractual, legal or regulatory purposes? Consider: o Research proposal (subsidieaanvraag) o Research protocol o Data/material transfer agreements o Intellectual property agreements o Informed consent forms o Data dictionaries Example answer: Research protocol (K:/Department/StudyA/Protocol/ StudyA_ResearchProtocol_Final_20180101.pdf). Informed consent forms: Digital version: K:/Department/StudyA/Forms/ InformedConsentform_ID1001.pdf; Paper version: Department of experimental immunology, AMC.	_____

8. INTELLECTUAL PROPERTY

8.1 Did you generate intellectual property?	◯ Yes ◯ No
8.3 Can you publish the existing dataset (as a part of the new, combined dataset) or is it protected by intellectual property?	_____

INITIAL PHASE QUESTIONS COMPLETED

Initital phase questions filled in? For the initial phase DMP all mandatory questions should be filled in. Choosing "Yes" will send an email to Durrer Center. Durrer Center will then assess the DMP and provide feedback within three weeks. Please note: this message will only be send the first time you select "Yes".	⊗ Yes

MID-TERM QUESTIONS COMPLETED

Mid-term questions filled in?

◯ Yes

For the mid-term DMP all questions should be filled in. Choosing "Yes" will send an email to Durrer Center. Durrer Center will then assess the DMP and provide feedback within three weeks.

Please note: this message will only be send the first time you select "Yes".