TANGO Data Management Plan Version 5

A Data Management Plan created using DMPonline

Creators: Valesca Retel, Jasmin K. Böhmer, Edwin Cuppen
Affiliation: Other
Template: ZonMw
Grant number: 846001002
Last modified: 14-12-2020
English Version: 14-12-2020

Table of Contents

1. GENERAL FEATURES OF THE PROJECT AND DATA COLLECTION

1.1 PROJECT LEADER CONTACT DETAILS

Valesca Retèl, PhD
Netherlands Cancer Institute
Division PSOE
Plesmanlaan 121
1066 CX Amsterdam
The Netherlands
tel: +31 20 512 6197
email: v.retel@nki.nl

1.2 I HAVE COMPOSED MY DMP WITH THE ASSISTANCE OF A DATA MANAGEMENT EXPERT.

The expert is attached to my department/institute.

- The expert is from outside my department/institute.

Jasmin Böhmer - Data Steward Utrecht Bioinformatics Expertise Core
Center for Molecular Medicine, University Medical Center Utrecht
+31 88 75 680 82
j.k.bohmer@umcutrecht.nl
bec@umcutrecht.nl

1.3 IN COLLECTING DATA FOR MY PROJECT, I WILL DO THE FOLLOWING:

- Using existing data (name)
- Generate new data
- Link data files (name)

Existing data:

- Genetic data: Whole Genome Sequencing (WGS) Data from the CPCT-02 study will be used and provided through the Hartwig Medical Foundation (HMF). Data requests have been made via the Data Access Board of the Center for Personalized Cancer Treatment (CPCT)/HMF and have been honored. The contract was concluded between HMF and Erasmus, HMF and Amsterdam VUmc, HMF and Amsterdam AMC.
- Clinical data: Clinical data of the CPCT-02 patients have been provided by selected (high volume) CPCT-02 hospitals. EMC, Meander, NKI-AvL, UMC Utrecht, AmsterdamUMC
- Registry data: in WP3 Santeon database is used for NSCLC and DMTR database is used for advanced melanoma, and IKNL data is used for modelling time-to-treatment patterns between hospitals (confirmed and received),
- Diagnostic pathway data: from the NKI-AVL, diagnostic pathway data has been used in NSCLC patients

New data:

- Analysis method for discovery and validation of biomarker for selection of non-responders to immunotherapy based on Whole Genome Sequencing (WGS) for advanced NSCLC and melanoma patients.
- Microcosting analysis to compare standard diagnostics to WGS
- Models of WP3,4,5
- Results of WP6 (ethical legal), reports on ethical/legal frameworks
- Questionnaires on quality of life, utilities and productivity of patients included in the CPCT-02 study. TANGO will record this in its own eCRF, the data remains property of NKI-AVL. After finishing the TANGO study, data can be requested to the TANGO project leader (V. Retèl).

Link data files:

- WGS result will be linked in the Pathologisch Anatomisch Landelijk Geautomatiseerd Archief (Palga)
- Clinical data from CPCT-02 study will be linked with the HRQOL-questionnaires
- WGS data will be linked to clinical data of the CPCT-02 patients of the hospitals concerning EMC, Meander, NKI-AvL, UMC Utrecht, AmsterdamUMC

1.4 I WILL DO HUMAN-RELATED RESEARCH.

- Yes

- Questionnaires on quality of life, utilities and productivity of patients included in the CPCT-02 study
- (Under the umbrella of the CPCT-02 human bound data is collected. We use the informed consent of the CPCT-02 study for the WGS result and clinical data.)
- For the questionnaires we use CPCT-02 logistics and TANGO is responsible for the questionnaires and takes care of the handling of privacy data in accordance with GDPR and manages the questionnaires and ensures their distribution. Patient can indicate in IC of CPCT-02 whether he/she wants to participate in the questionnaires. After finishing the TANGO study, data can be requested to the TANGO project leader (V. Retèl).

1.5 I WILL REUSE EXISTING DATA AND I HAVE PERMISSION FROM THE DATA OWNER / OWNERS FOR THE USE OF HIS / THEIR DATA.

- Yes, I have permission to use their data
- Yes, I have permission to use their data, but I have to destroy it after the end of the project.

Clinical & Genetic:

1) *Whole Genome Sequencing Data (WGS)*: HMF data was approved and provided to EMC; VU/AMC

2) *Clinical Data*: the contractual situation between the EMC and their partnering hospitals is clear; the contractual situation between VU and their partnering hospitals is still under way

3) *Patient Data Registries*: the contracts between VU and DMTR, and Santeon are signed and the data transfer approved; the contract between VU and NVALT is still in application status

4) *Costing Data*: The contracts between UMCU and Rijnstate, and NKI are in place and the data is transferred; the contracts between UT and NKI, AND UMCU, and Rijnstate are in place and the data is transferred.

5) *Quality of Life (QoL):* the contractual situation between NKI and UMCU/CPCT is still under review; the contractual situation between NKI and CPCT for clinical data is still in review status.

1.6 I WILL LINK EXISTING DATA AND I HAVE MADE AGREEMENTS WITH THE DATA OWNER / OWNERS FOR THE LINK.

- Yes

It is intended that the linking of WGS results will be organised in this project with PALGA (https://www.palga.nl/). A declaration of intent has been signed by HMF (https://www.hartwigmedicalfoundation.nl/) and PALGA for this purpose. The linking of the clinical data and WGS data of the CPCT-02 (https://www.cpct.nl/cpct-02/) patients has been arranged and falls within the CPCT-02 Informed Consent.

1.7 WHEN COLLECTING DATA, I WORK TOGETHER WITH OTHER PARTIES.

- Yes, the data to be collected are (partly) supplied by a partner of the project or supplier.

- Yes, I collect research data together with other researchers, research groups.

  - WGS data is supplied by HMF/CPCT-02.
  - Collection and generation of research data will take place within a Consortium, consisting of NKI, UMC Utrecht, University of Twente, Erasmus MC, MUMC+ and AmsterdamUMC.
  - Because the research takes place within a Consortium, a Consortium Agreement has been concluded between the participating centres.

1.8 I FORESEE THE FOLLOWING END PRODUCTS OF THE PROJECT AND MAKE THESE AVAILABLE FOR FOLLOW-UP RESEARCH AND VERIFICATION.

- (different versions of) Processed data

- Documentation about the data

- Documentation of the research process, including all parties involved

Raw sequencing data cannot be made available because it is the responsibility of HMF. This data has been requested via the HMF Data Access Board with a formal request (more information about the procedure and application forms can be found at www.hartwigmedicalfoundation.nl). HMF's bioinformatic pipeline is based on open-source software and publicly available at GitHub (https://github.com/hartwigmedical/pipeline).
Explanation of the end products that we provide and that can be made available:

- Processed data: subset of WGS results

- Documentation: code book on methodology
- Protocols
- Possibly software related to models (www.anylogic.com)

1.9 I CAN MAKE AN ESTIMATE OF THE SIZE OF THE DATA COLLECTION

namely the number of participants or subjects ("n =") of the data collection and the size in gigabytes / terabytes.

- Yes (name)

n = 400 NSCLC patients
The raw data and analysis files of the WGS data are managed by HMF. These can be viewed remotely if necessary, which does not make it necessary to copy many terabytes of data. The gVCF files are used for the actual analysis and must be stored within TANGO (at least until the analysis is completed). Original files are safe with HMF in an ISO/NEN7510 accredited private cloud (Schuberg-Phillis).

Calculation:

- 20Gb per patient for WGS analysis (gVCF) = 8TB
- Models: 5-10MB per model, about 5Gb in total
- Questionnaires in Word, analysis in SPSS, max 1 Gb (?)

Total footprint:

- 8TB + 5Gb + 1Gb ~> 8TB

Institution Acronym	Work package	Overall Type of Data	Overall Estimated Data-Size
AMC	WP1	Whole Genome Sequencing; Molecular Data	~1TB – 2TB
UMCU	WP1	Healthcare Cost and Consumption Data	<10GB
EMC	WP2	Whole Genome Sequencing	~1-1.5TB
VUmc	WP2	Whole Genome Sequencing	~1-1.5TB
VUmc	WP3	Patient Registry data	< 10GB
MUMC+	WP4	Survey Data; Patient Registry Data	<100GB
UT	WP5	Survey Data; Patient Registry Data; Simulation Data	<100GB
AMC/UMCU	WP6	Survey Data; Interview data	<100GB

WP1 - WGS and NGS Data				
Data Type	Data File Format	Required Software	Estimated Data Size	Special Characteristics
WGS	VCF	*tbd*	~125GB per patient	Coded and pseudonymised
Patient Records and Clinical Diagnostics data	Proprietary Patient Record File	*tbd*	<100GB	Not anonymised
Next Generation Sequencing	VCF	*tbd*	NGS whole exome VCF 1GB	Not anonymised
Gene Panels	VCF	*tbd*	NGS gene panel (50 genes) 1MB, NGS gene panel (500 genes): 10MB	Not anonymised
Immunohistochemical Data	TIFF	*tbd*	<100GB	Not anonymised
Output Data	Graphical Illustrations	Microsoft Office Suite	<100GB	Publication related content

WP1 - Healthcare Cost and Consumption Data				
Data Type	Data File Format	Required Software	Estimated Data Size	Special Characteristics
Tabular Data	CSV; Microsoft Excel	Microsoft Excel	<10GB	Coded
R-Scripts	R	R, R-Studio	<10GB	Coded

WP2 - WGS and Clinical Data EMC				
Data Type	Data File Format	Required Software	Estimated Data Size	Special Characteristics
Clinical Data	TXT; Microsoft Excel	Microsoft Excel	<1GB	CPCT-02/ HMF identification number included
WGS raw data	BAM	R Studio	~1.2TB	CPCT-02/ HMF identification number included
WGS variants	VCF	IGV	<5GB	Not anonymised

WP2 - WGS and Clinical Data VUmc				
Data Type	Data File Format	Required Software	Estimated Data Size	Special Characteristics
Clinical Data	TXT; Microsoft Excel	Microsoft Excel	<1GB	CPCT-02/ HMF identification number included
WGS raw data	BAM	R Studio	~1.2TB	CPCT-02/ HMF identification number included
WGS variants	VCF	IGV	<5GB	Not anonymised

WP3 – Analysis of survival pattern Data				
Data Type	Data File Format	Required Software	Estimated Data Size	Special Characteristics
Data type	Data File Format	Requirement Software	Estimated Data Size	Special Characteristics
Patient Registry Data	sav: SPSS csv: Microsoft Excel	SPSS Microsoft Excel R	< 10 GB	Pseudonymised

WP4 - Cost-effectiveness Analysis Data				
Data Type	Data File Format	Required Software	Estimated Data Size	Special Characteristics
Data registries	SPSS; CSV; R-file	SPSS; Microsoft Excel; R statistics	<100 GB	pseudonymised
Survey data	SPSS; CSV; R-file	SPSS; Microsoft Excel; R statistics	<100 GB	No personal data included
Simulation data	CSV; R-file	Microsoft Excel; R statistics	<100 GB	*n.a.*

WP5 – WGS Implementation Requirements Analysis				
Data Type	Data File Format	Required Software	Estimated Data Size	Special Characteristics
Survey Data	CSV	Microsoft Excel	<100 MB	*n.a.*
Patient Registries	XLS; CSV	Microsoft Excel	<100 MB	Confidential, pseudoanonymized
Simulation Data	CSV	Microsoft Excel	< 1GB	*n.a.*
Cost Data	XLS; CSV	Microsoft Excel	<100 MB	*n.a.*
Simulation Model	ALP	AnyLogic	<100GB	*n.a.*

1.10 DURING THE PROJECT I HAVE SUFFICIENT STORAGE LOCATIONS AND CAPACITY AND I HAVE A BACK-UP OF THE DATA AVAILABLE.

- Yes, for the storage and backup of my data I use the standard facilities of my institute.

  - WGS data: raw and analysis data is stored at HMF, is safely stored at Schuberg Phillis, of which a mirror is also available.
  - Each institute is responsible for the secure storage of derived data, analysis protocols and software on its own servers. They must also provide a regular backup (this will be included in the terms of use). Utrecht Bio-informatics Expertise Centre (UBEC) will include this as an extra step in the periodic DMP check (which is done every six months). So besides re-evaluating the DMP and correcting it with the state of affairs at that moment, we will contact the participants beforehand to see if and how the conditions of use are met.
  - For the HMF, the responsibility for auditing the DMP lies with the Data Access Committee/Board of HMF/CPCT, which has been accredited since the beginning of 2018.

In November 2019 the PhD students and junior researcher have received a comprehensive training session about research data management, in which the ins and outs of a solid back-up routine was taught.

2. LEGISLATION (INCLUDING PRIVACY)

2.1 I WILL BE DOING RESEARCH INVOLVING HUMAN SUBJECTS

, and I am aware of and compliant with laws and regulations concerning privacy sensitive data.

- "Wet Bescherming Persoonsgegevens" and the resulting code of "gedragscode Gezondheidsonderzoek". I register my project with the "Autoriteit Persoonsgegevens"
- Wet Medisch-Wetenschappelijk Onderzoek with people (WMO)
- Wet op de Geneeskundige Behandelingsovereenkomst

- Aware of the WBP, but we do not need to report personal data to the "Autoriteit Persoonsgegevens", we will report this to the Data Protection Officer (FG) of the UMC Utrecht. After reporting, a PIA was also drawn up in collaboration with the FG. This will be evaluated annually, after evaluation of the DMP.
- CPCT-02 is WMO mandatory research, METC approval obtained at UMCU, permission has also been applied for locally and granted at centres participating in the CPCT research. Access to the data is subject to certain conditions.
- Linking the clinical data to the CPCT-02 data may be possible using the CPCT number. In practice, however, this is not possible (data access boards do not give permission to link the data to other sources). This appears to be a legal obstacle in particular, as there is no clear correspondence about the legislation. A solution that is still being explored is the retrieval of the data (what data?) per hospital via the PI of the CPCT. At this moment (April 2018) it is being investigated whether the analysis can take place per hospital dataset, or whether it is necessary to combine the data. One option may then be that the hospital itself makes the link and then safely stores or deletes the code. The data can then be transported to the person responsible within TANGO.
- Generating data from questionnaires will most likely be done via non-WMO mandatory research.
- WGBO applies to lung doctors.
- There is a consortium agreement in place. Additionally, every individual UMC/University puts contracts with their partnering institutions in place complying to their institutional legislation and contractual requirements.

2.2 I WILL BE DOING RESEARCH INVOLVING HUMAN SUBJECTS,

and I have (a form of) informed consent from the participants for collecting their data.

- Yes, please specify the form of consent.
- Yes, the form of consent allows reuse of the data (further use).
    - Via informed consent of the CPCT-02 study.
    - Within the informed consent of the CPCT-02, permission is requested for sending questionnaires. The amendment currently lies with the METC of the UMCU. To link the questionnaires and clinical data, formal data request is submitted to HMF/ CPCT-02.

2.3 I WILL BE DOING RESEARCH INVOLVING HUMAN SUBJECTS,

and I will protect my data against misuse.

- Yes, I have the data pseudonymized.

Data released shall be pseudoanonymised. The study ID (code) is stored separately, and is very limited accessible (management of key is done by a data manager at one of the centres).
Patients have a CPCT -02 ID number, which is registered in the EPD. This CPCT-02 ID number will also be used for the questionnaires.

2.4 I WILL STICK TO THE PRIVACY REGULATIONS OF MY ORGANISATION

- Yes

All participants in the study are working within UMCs with corresponding privacy regulations for all employees.

3. MAKING DATA FINDABLE

3.1 THE DATA COLLECTION OF MY PROJECT WILL BE FINDABLE FOR SUBSEQUENT RESEARCH

- Yes, via the search engine of the archive (repository) in which it is stored (name)

Zenodo was chosen as dedicated long-term archive for most of the outputs from this project. A community was created on Zenodo: https://zenodo.org/communities/tango-wgs/

Other website that refer to the archival collection are: the ZonMW project website [https://www.zonmw.nl/nl/onderzoek-resultaten/geneesmiddelen/programmas/project-detail/personalised-medicine/technology-assessment-of-next-generation-sequencing-in-personalized-oncology-tango/], the CPCT-02 study website [https://www.cpct.nl/cpct-02/], and the project record on NARCIS [https://www.narcis.nl/research/RecordID/OND1361542].

Up until now, mainly presentation and poster publications are provided on the Zenodo archive.

Raw data: The received data from the CPCT study, or additional data requested from hospitals remains with the providing institutions. For reuse-purposes the Data Access Committees of each individual institution can be contacted. That means that the raw data remains under restricted access.
Anonymised raw data from the non-genomic data will be made accessible as open as possible, as closed as necessary.

Processed data: the selection of the appropriate repositories and certified archives is still open. Restricted access is appropriate due to the sensitive content of the data-files. Processed data related to non-genomic data will be made accessible as open as possible, as closed as necessary.

Output data: non-sensitive data will be made openly accessible as possible. Open Data is desired.

Available access restrictions based on 4.4:

CODE	RESTRICTION
RQ	Limitations with regard to the research questions to be answered.
CO-OP	Cooperation in the use of the dataset, including agreements on publications, authorships.
METH	Agreements on methodology.
LINK	If the dataset may be linked to another dataset (privacy).
COMM	Sharing data for commercial purposes. In doing so, I take into account the provisions of state aid law
SEC	Conditions relating to data security.
ACCESS	The way in which the dataset is made available.
PERIOD	The period of permission for use of the dataset.
REIMB	The reimbursement of costs, for example for obtaining research data.
COM	A steering committee, program committee or project leader will decide on the approval of data applications.
OTHER	define other access restrictions yourself
OA	this is an open access item

Institution Acronym	Work package	Overall Type of Data	Overall Estimated Data-Size for Archiving
AMC	WP1	Whole Genome Sequencing; Molecular Data	
UMCU	WP1	Healthcare Cost and Consumption Data	
EMC	WP2	Whole Genome Sequencing	
VUmc	WP2	Whole Genome Sequencing	<100 GB
VUmc	WP3	Patient Registry data	<100 GB
MUMC+	WP4	Survey Data; Patient Registry Data	<100 GB
UT	WP5	Survey Data; Patient Registry Data; Simulation Data	<10 GB
AMC/UMCU	WP6	Survey Data; Interview data	

WP1 - Implementation of WGS in the Clinical Routine – Publishing and Archiving								
Data Stage	Data Type	Preservation File Format	Recommended Software	Estimated Data Size	Special Characteristics	Extra documentation	Access Restriction	Contact
RAW	Gene Mutation data from patient registry	.xlsx	spss/R/Excel	< 1GB	CPCT-02 HMF identification number	Readme file	CO-OP	na
	Minimal Patient Registry data	.xlsx	spss/R/Excel	< 1GB	CPCT-02 HMF identification number	Readme file	CO-OP	na
	Gene Mutation data from HMF WGS Databse	.xlsx	spss/R/Excel	< 1GB	CPCT-02 HMF identification number	Readme file	CO-OP	na
	Survey data Molecular Tumor Boards	.xlsx	spss/ /Excel	< 1GB	none	Readme file	CO-OP	na
PROCESSED	Merged dataset from mutational data patient registry and derived from	.xlsx	spss/ /Excel	< 1GB	CPCT-02 HMF identification number	Readme file	CO-OP	na

	HMF databse							
OUTPUT	Graphs, tables, text	.xlsx .pdf	common	< 1GB	none	none	OA	na

WP1 – Healthcare Cost and Consumption Data – Publishing and Archiving								
Data Stage	Data Type	Preservation File Format	Recommended Software	Estimated Data Size	Special Characteristics	Extra documentation	Access Restriction	Contact
RAW	*Still under review. Will be updated in next DMP version.*							
PROCESSED	*Still under review. Will be updated in next DMP version.*							
OUTPUT	*Still under review. Will be updated in next DMP version.*							

WP2 - WGS and Clinical Data EMC – Publishing and Archiving								
Data Stage	Data Type	Preservation File Format	Recommended Software	Estimated Data Size	Special Characteristics	Extra documentation	Access Restriction	Contact
RAW	WGS data	.bam	R Studio Python	1.2 TB	CPCT-02 HMF identification number	none	CO-OP	HMF requested data (DR-008). j.aerts@erasmusmc.nl
	WGS variant	.vcf	IGB	< 5GB	CPCT-02 HMF identification number	none	CO-OP	HMF requested data (DR-008). j.aerts@erasmusmc.nl
	Clinical data CLINT database Erasmusc MC	.txt .xlsx	Common	< 1GB	Pseudonimized	README_ WP2_Clin.txt	CO-OP	j.aerts@erasmusmc.nl
PROCESSED	Codes for analysis WGS	.py	Python	< 1GB	none	none	COM	e.voest@nki.nl
	Codes for analysis clinical data	.r	R	< 1GB	none	none	COM	j.aerts@erasmusmc.nl
OUTPUT	Graphs and tables	.pdf .ai	Common	<10GB	none	none	COM	j.aerts@erasmusmc.nl

WP2 - WGS and Clinical Data VUmc – Publishing and Archiving								
Data Stage	Data Type	Preservation File Format	Recommended Software	Estimated Data Size	Special Characteristics	Extra documentation	Access Restriction	Contact
RAW	Clinical data	.sav	SPSS	< 1GB	Pseudonimized	none	CO-OP	vande-neertwegh@amsterdamumc.nl
	WGS data	.bam	R Studio	1.2 TB	CPCT-02 HMF identification number	none	CO-OP	HMF requested data (DR-079). vande-neertwegh@amsterdamumc.nl
PROCESSED	codes for analysis clinical data	.r	R Studio	< 1GB		none	COM	vande-neertwegh@amsterdamumc.nl
	codes for analysis WGS	.py	Python	< 1GB		none	COM	e.voest@nki.nl
OUTPUT	Graphs and tables	.pdf .ai	Common	<10GB	none	none	COM	vande-neertwegh@amsterdamumc.nl

WP3 – Analysis of survival pattern Data – Publishing and Archiving								
Data Stage	Data Type	Preservation File Format	Recommended Software	Estimated Data Size	Special Characteristics	Extra documentation	Access Restriction	Contact
RAW	Santeon lung cancer registry 2008-2014	.sav	SPSS	164 KB	none	none	CO-OP	Santeon Onderzoek
	DMTR-registry 2017	.csv	common	18.4 MB	none	none	CO-OP	Dutch Institute for Clinical Audit
	NKI lung cancer tumour volume	.csv	common	10 MB	none	none	CO-OP	NKI – Immuno-Radiology
PROCESSED	Simulated data	.csv	R	<10 GB	none	none	OA	
	Code for model building and analysis	.r	R	<10 GB	none	none	OA	
OUTPUT	Extra output on model validation (not reported	.pdf	Common	<10 GB	None	None	OA	

	in main article)							

WP4 - Cost-effectiveness Analysis Data – Publishing and Archiving								
Data Stage	Data Type	Preservation File Format	Recommended Software	Estimated Data Size	Special Characteristics	Extra documentation	Access Restriction	Contact
Raw	Patient registry	proprietary	Bizzmine (online platform)	<100 GB	pseudonymized	none	CO-OP	HMF data: v.retel@nki.nl
	Patient questionnaire	.csv .xlsx	common	<1GB	pseudonymized	none	CO-OP	TANGO data: v.retel@nki.nl
OUTPUT	Digitized patient level time to event data	.csv	common	<1GB	none	DOI: 10.5281/zenodo.4288605	OA	i) MUMC+ / martijn.simons@mumc.nl ii) MUMC+/ m.joore@mumc.nl iii) NKI-AvL / v.retel@nki.nl
	Survey data	.csv .xlsx	common	<1GB	none	none	OA	i) MUMC+ / martijn.simons@mumc.nl ii) Utwente / m.vandeven@utwente.nl iii) NKI-AvL / v.retel@nki.nl

WP5 – WGS Implementation Requirements Analysis – Publishing and Archiving								
Data Stage	Data Type	Preservation File Format	Recommended Software	Estimated Data Size	Special Characteristics	Extra documentation	Access Restriction	Contact
RAW	*Still under review. Will be updated in next DMP version.*							
PROCESSED	*Still under review. Will be updated in next DMP version.*							
OUTPUT	Survey data	.csv .xlsx	common	<1GB	none	none	OA	i) MUMC+ / martijn.simons@mumc.nl ii) Utwente / m.vandeven@utwente.nl iii) NKI-AvL / v.retel@nki.nl

3.2 I WILL USE A METADATA SCHEME FOR THE DESCRIPTION OF MY DATA COLLECTION.

- Yes, I choose a generic metadata schema (called)

The basic metadata standard Dublin Core will be applied, which is compliant to Datacite 4.0. This metadata standard is applied to the whole data-set and will enable the findability of the research data.

3.3 I WILL BE USING A PERSISTENT IDENTIFIER AS A PERMANENT LINK TO MY DATA COLLECTION

- Yes, the doi-code

Update December 2020: The Zenodo community of this project is equipped with a URL, however all published datasets via Zenodo receive a DOI.

4. MAKING DATA ACCESSIBLE

4.1 ONCE THE PROJECT HAS ENDED, MY DATA WILL BE ACCESSIBLE FOR FURTHER RESEARCH AND VERIFICATION.

- Yes, after an embargo period (explain).

The data will be made available after an embargo period of 3 months after publication of the results generated on the different datasets.

There are several reasons for this:
- Many different parties
- Multiple publications
- Longer contract period of individual researchers

The data from a number of sub-projects can be expected to be made available within the foreseeable future, for example the cost analysis of WGS versus standard diagnostics.
We aim to make the data available as soon as the relevant sub-component has been completed and published.
The WGS data is not generated in this study itself, but is accessible via the Hartwig Medical Foundation.

Please refer to 3.1 to see where what type of output was stored/archived and who to contact.

4.2 ONCE THE PROJECT HAS ENDED, MY DATA COLLECTION WILL BE PUBLICLY ACCESSIBLE, WITHOUT ANY RESTRICTIONS (OPEN ACCESS).

- No, I attach conditions to access to the data collection (restricted access) (explain).

Conditions are linked to access to the data collection for the following reasons:
- Part of the data is personally identifiable and therefore privacy sensitive.
- Part of the data is already available on request under certain conditions (for example, the raw data of HMF can be retrieved via data access board request).

If the above restrictions do not apply to a certain data type (e.g. static information at an aggregated level) and it is allowed according to informed consent, this data will be made publicly available.

Update December 2020:
The Zenodo community of this project features all open access publications, the DMP will list all dedicated long-term storage locations and contact details for restricted and closed access outputs.
Please refer to 3.1 to see where what type of output was stored/archived and who to contact.

4.3 I HAVE A SET OF TERMS OF USE AVAILABLE TO ME

, which I will use to define the requirements of access to my data collection once the project has ended (please provide a link or persistent identifier; also note that this is a key item, which you should report to ZonMw at the conclusion of your project).

- No, my institute is going to draw up the terms of use in cooperation with a lawyer.

Depending on the dedicated data archive and its license agreement, a mix of licenses is anticipated. In case the license options of the data archive are not sufficient or not available, license files will be provided with each data-set.

Update December 2020:
Please refer to 3.1 to see where what type of output was stored/archived and who to contact.
All outputs available on the Zenodo-community are equipped with the appropriate license; with CC-BY 4.0 as default license.

4.4 IN THE TERMS OF USE RESTRICTING ACCESS TO MY DATA

, I have included at least the following:

- Limitations with regard to the research questions to be answered.
- Cooperation in the use of the dataset, including agreements on publications, authorships.
- Agreements on methodology.
- If the dataset may be linked to another dataset (privacy).
- Sharing data for commercial purposes. In doing so, I take into account the provisions of state aid law
- Conditions relating to data security.
- The way in which the dataset is made available.
- The period of permission for use of the dataset.
- The reimbursement of costs, for example for obtaining research data.
- A steering committee, programme committee or project leader will decide on the approval of data applications.

We must adhere to the informed consent of the CPCT-02 study, so we cannot and will not share the data just like that.

Update December 2020:
Please refer to 3.1 to see where what type of output was stored/archived and who to contact.

5. MAKING DATA INTEROPERABLE

5.1 I WILL SELECT A MACHINE ACTIONABLE DATA FORMAT

, which will allow other researchers and their computers to read my data collection.

- Yes

So far known:
Excel: in CSV format
Word/R/C++: in plain text, if possible, within a version control system such as Git/SVN
SPSS: in CSV format

Data is processed in Word and Excel by the researchers themselves, but distributions in text and CSV files will be delivered, which are interoperable. The linked metadata will make the data partly machine readable. A pilot project is underway to deliver datasets as FAIR data points (including FAIR distribution of the dataset). If the pilot is completed with good results, this project would be eligible for FAIR conversion.
The sequencing data is stored in FASTQ format at HMF. The analyses are done on VCF format. Both are a standard in bioinformatics, both readable via a text editor.'.

Update December 2020:
Please refer to 3.1 to see where what type of output was stored/archived in which file-format and who to contact.

5.2 I WILL SELECT A METADATA STANDARD TO ALLOW MY DATA COLLECTION

to be linked to other collections (note: this is a key item, which you should report to ZonMw at the conclusion of your project).

- Yes, I choose a metadata standard from the overview of Biosharing

The basic metadata standard Dublin Core will be applied, which is compliant to Datacite 4.0. Where applicable, domain specific metadata will be captured and provided in addition to Dublin Core. This metadata standard is applied to the data file level to enable the interoperability of all the files within the data-sets.

5.3 I WILL BE DOING RESEARCH INVOLVING HUMAN SUBJECTS, AND I HAVE TAKEN INTO ACCOUNT THE REUSE OF DATA

and the potential combination with other data sets when taking privacy protection measurements.

- Yes, the participants have given permission for further use of the data and the data have been pseudonymised.

See previous answers.

Update December 2020:
We will list the reused/requested data from the hospitals etc, and note that the consent is available and everything is covered by the individual DTA/MTA. The project coordinator has an overview of all available legal documents and documentation; these information are not shared in this public DMP.

6. MAKING DATA REUSABLE

6.1 I WILL ENSURE THAT THE DATA AND THEIR DOCUMENTATION WILL BE OF SUFFICIENT QUALITY

to allow other researchers to interpret and reuse them (in a replication package).

- I document the research process (explain)
- I perform quality checks on the data so that they are complete, correct and consistent (explain)

Research process:
- Protocols and research proposal

Quality checks:
- SOPs and settings WGS equipment
- Check presence of informed consents
- SOPs about data cleaning

HMF works according to ISO17025 accreditation. Lab protocols (SOPs) are available on request for interested parties.

For bioinformatic data analysis, all pipeline and tools are publicly available and versions are maintained. These can be found at GitHub (https://github.com/hartwigmedical).

- A data privacy impact assessment (DPIA) was performed in July 2017 and was update in February 2020
- A research data management audit was performed in October 2018
- The data management plan was updated in December 2018, 2019, 2020

Update December 2020:
Please refer to 3.1 to see where what type of output was stored/archived in which file-format, equipped with what documentation, and who to contact.
Output underlying publications that have been published/archived on Zenodo are equipped with a standardised re-adme-file, and refer to the publication where the methodology is provided.

6.2 I HAVE A NUMBER OF SELECTION CRITERIA

, which will allow me to determine which part of the data should be preserved once the project has ended. (see also question 1.9)

- YES

- VCF data as obtained from HMF does not need to be stored locally after completion of the study, as these data can be requested from HMF.
- The rest of the data is still difficult to estimate.

It is intended to create an archived collection that is as open as possible and as closed as necessary, while providing as much of the relevant processed and output data.

Update December 2020:
Zenodo will feature the publication related data, posters, and slides. As well as any relevant auxiliary data/code.
Please refer to 3.1 to see where what type of output was stored/archived and who to contact.

6.3 ONCE THE PROJECT HAS ENDED AND THE DATA HAS BEEN SELECTED

, I can make an estimate of the size of the data collection (in GB/TB) to be preserved for long-term storage or archival.

- Yes(name)

Relates to 1.9, however since the original raw is supplied by the HMF and other medical centres, there is no need to archive them again. Therefore, the overall estimated data volume for archiving ranges between 200GB and 500GB.

Update December 2020
Currently, the estimated data-size of the whole project is <500GB.

6.4 I WILL SELECT AN ARCHIVE OR REPOSITORY FOR (CERTIFIED) LONG-TERM ARCHIVING

of my data collection once the project has ended. (note: this is a key item, which you should report to ZonMw at the conclusion of your project)

- YES

See question 3.1. The appropriate solution is still under consideration.
The data from the questionnaires will be included in the CPCT-02 dataset to keep the data accessible in the future.

Update December 2020:
The established Zenodo community of this project, and all its previous publications have led to the decision to make Zenodo the main archive for all outputs. The DMP will list all dedicated long-term storage locations and contact details for restricted and closed access outputs.

6.5 ONCE THE PROJECT HAS ENDED, I WILL UPHOLD THE RECOMMENDED DATA PRESERVATION PERIOD OF AT LEAST 10 YEARS.

- yes (state number of years)

Minimum 10 years. We will not retain the data longer than necessary for the purpose for which it was collected or to comply with legal or regulatory requirements. After this, the data will be deleted.
How is it ensured that all parties comply with these conditions?

Update December 2020:
Currently a final overview is in progress and will be p[provided in the next version of this DMP.

6.6 DATA MANAGEMENT COSTS DURING THE PROJECT

and preparations for archival can be included in the project budget. These costs are:

- Yes (Explain)

In the budget 10% is reserved for data stewardship.