IMPALA DMP

A Data Management Plan created using DMPonline

Creator: María José Villalobos Quesada

Affiliation: Leiden University Medical Center

Funder: European Commission

Template: LUMC data management plan

ORCID iD: 0000-0001-6522-1347

Project abstract:

Background:

More than 3 million children in low-resource settings (LRS) die annually due to contextual constraints in healthcare systems that hamper the widespread supply of high-quality healthcare. Many of these deaths are advanced stages of poverty-related diseases that are recognised too late to be treated effectively while treatment usually is available.

In the hospital, monitoring of children's vital signs is essential for early detection of critical illness, which can save lives. Shortage of (qualified) staff and lack of suitable equipment are the main bottlenecks to monitor these children adequately during admission. Current monitoring systems widely used in high-resource settings are not suitable for LRS due to their high costs and poor compatibility with LRS settings.

The IMPALA project will address these problems by developing an affordable, durable, and user-friendly monitoring system (IMPALA) for hospitalised children in LRS. By combining innovative sensors, machine learning algorithms and point-of-care biomarkers we aim to create a smart, yet simple, monitoring system that enables health workers to timely detect and predict critical illness. At the end of this project, we aim to have a fully functional monitoring system that will be ready to be tested in clinical settings.

General objective:

To further develop a current tablet-assisted monitoring system that is suitable for paediatric-hospital care in LRS which will include the real-time prediction of critical events through predictive algorithms based on vital signs (optionally supplemented with clinical data and biomarkers), allowing timely and lifesaving interventions.

Specific objectives:

1. To conduct extensive implementation research to identify key barriers and opportunities to implementation of vital signs monitoring in LRS, and to address these barriers appropriately.
2. To develop algorithms that predict critical illness based on vital signs and to enhance their accuracy by also incorporating clinical data and/or biomarkers.
3. To design the future randomised clinical trial to assess the impact of the monitoring device and the implementation strategy on in-hospital paediatric survival in LRS.

ID: 83014

Start date: 01-06-2021

End date: 31-05-2025

Last modified: 12-01-2022

Grant number / URL: https://www.edctp.org/edctp2-project-portal/

IMPALA DMP

I. General information

I.1 Name of researcher responsible for DMP

Dr. María Villalobos

I.2 Department of researcher responsible for DMP

PHEG

I.3 ORCID ID of researcher responsible for DMP

https://orcid.org/0000-0003-4930-1982

I.4 Supervisor(s) of project, if applicable

Project's principal investigator is **Dr. Job Calis**, Stichting Amsterdam Institute for Global Health and Development (AIGHD). The project has been design in five hubs according to the nature of the research activities and the data generated and processed.

Biological samples and biomedical research will be under the responsibility of **Dr. Myrsini Kaforou*** (Imperial College London).

Dynamic vital signs research and IMPALA monitoring system will be under the responsibility of **MSc. Bart Bierling*** (GOAL 3).

Clinical research and data will be under the responsibility of **Dr. Jenala Njirammadzi*** (KUHeS).

Social sciences and implementation research will be under the responsibility of **Prof. Wendy Janssens*** (AIGHD).

Data management and data sciences research will be under the responsibility of **MSc. William Nkhono*** (TRUE).

*The leaders of each hub will be defined in the next Management Team Meeting in January 2022. This is the last proposal made by the Management Team end 2021.

I.5 If applicable for your study or project, please provide:

If one or more numbers are not applicable for your study or project, please add '/ ' in the appropriate text box.

EDCTP-sponsored Pan African Clinical Trials Network and on Clinicaltrials.gov	To be defined shortly.
METC number	Project will be submitted to COMREC, the METC of Kamuzu University of Health Sciences (KUHeS).

I.6 If applicable: list the partner organisation(s) involved in your study or project.

Kamuzu University of Health Sciences (KUHeS), Malawi

Malawi University of Business and Applied Sciences (MUBAS-The Polytechnic), Malawi

Training and Research Unit of Excellence (TRUE), Malawi
Imperial College London, UK
Stichting Amsterdam Institute for Global Health and Development (AIGHD), the Netherlands
Vrije Universiteit Amsterdam
GOAL 3 BV, the Netherlands
National eHealth Living Lab (NeLL) LUMC, the Netherlands

II. About this DMP

II.1 Date of first DMP version

2021-10-28

II.2 Consulted LUMC data stewardship expert(s)

Note: This field is a requirement for most funders. DMPs will only be reviewed by funders after you have consulted one of the LUMC data stewardship experts.

To request for review: see section 7.

Question not answered.

II.3 Changes made to an earlier version of this DMP

28-10-2021 first version was drafted.
12-01-22 the draft was finished. Changes will be recorded after this point.

Part of DMP	Date of change (in dd/mm/yyyy)	Question number(s)	Adaptation(s) made
I. General information			
II. About this DMP			
1. Data collection			
2. Data documentation			
3. Data storage and security			
4. Data access, sharing and reuse			
5. Data preservation and archiving			
6. Additional information			

1. Data collection

1.1 What type of study or project will you conduct?

Use the 'additional information' field to briefly describe the data assets you will produce during the research process.

- Study/project combining human participants and human material(s)

The IMPALA project is a clinical trial which will recruit human participants and collect different types of primary data generated and/or processed:

I. Non-clinical data will be generated during the **implementation studies** (including social science and usability and user-centred design research using mix methods and participative approaches) and will collect end-user data for improving and evaluating the monitoring system in two sites in Malawi (WP3 and WP4).

End-user data includes:

- voice and video recordings generated during interviews, focus groups and personal logs which will be transcribed or summarised,
- ethnographic/contextual inquiry research notes,
- Electronic case report forms (eCRfs) questionnaires designed using digital tools (such as KOBO toolbox), if not possible they will be done on paper and then digitalised by a researcher, the original will be destroyed once digitalised.
- usage data (recordings of the use of the user interface in terms of clicks, and interaction with the interface),
- system data (e.g. errors, battery, connection, signals and alarm logs)
- costing data (e.g. of medicines, diagnostic tests, (average) salaries, bed occupancy),
- drug inventory data.
- observation data which will be summarized

II. Clinical data includes:

- data from the patient health records (analogue) that will be manually entered in an eCRF (designed using REDCap),
- clinical data reported directly into the eCRF,
- vital signs data generated by the sensors attached to patients,
- RNA and protein signatures (these are small signatures, no high-throughput data is foreseen),
- biological samples such as blood and swabs will be collected from patients.

The clinical data will be used for the training of a machine learning algorithm, and it will be generated during a medium-scale clinical observational study in two sites in Malawi: Zomba Central Hospital (Zomba) and Queen Elizabeth Central Hospital (Blantyre). Additionally clinical data will be used for developing and improving the monitoring system, as well as for assessing the effect (pre-post) on clinic practices and patient outcomes of introducing the monitoring system compared to the status quo.

Clinical data includes medical history, laboratory results, treatments and medication, vital signs and other medical data recorded during the stay of patients in the hospital, and genetic (RNA) and proteomic data. These two last ones will be isolated from blood samples obtained at admission.

The goal of the project is to generate a final prototype of the IMPALA monitoring system, developed in and for low resource settings, that will be fully functional and wholly co-developed with end-users, ready for a future large-scale validation in a multi-centre, international Randomised Clinical Trial (RCT), which is **not** currently part of this project.

1.2 Describe what you will do to pseudonymize or anonymize your data: How will you pseudonymize, where will identifiable data be stored and who is responsible for managing this data during the study or project?

Informed consent will be stored onsite where it was collected (Zomba or Queen Elizabeth Central Hospitals) in a safe, double locking location already designated for this purpose.

Clinical data will be **pseudonymised** by flagging and excluding any highly identifiable information of

the eCRF during data exports (REDCap will be used to design the eCRF). One measure taken will be to assign random research numbers to each participant; this number will not contain any personal-related information, only the location where it was generated (Zomba or Queen Elizabeth Central Hospital). A master file key in MS docx format will be generated and stored where data is generated, specifically at Queen Elizabeth Central Hospital, or Zomba Central Hospital, in secure servers with automatic backup and/or secure double locked location in paper is used... key will be necessary to link the data contained in the eCRF to the monitoring data and the patient. The data managers at TRUE will have control over the key and will facilitate access to others when considered necessary (e.g. the project's data manager, for the project's monitoring purposes or upon the request of authorities).

Non-clinical data obtained during the implementation research will also be pseudonymised by removing identifying information from the transcripts and replacing it with a random research number similar to that assigned with the clinical data to enable linkage between the two (2) datasets. Other notes/observations and voice/video recordings taken by researchers will be processed into text, saved as de-identified MS docx files and the original handwritten notes/files will be archived and backuped to a secure encrypted server/backup. These data will be made available when considered necessary during and/or at the end of data analysis and dissemination to allow for additional validation and quality control both internally and with monitors.

Another privacy-preserving strategy that we will apply is **data minimalisation** (e.g. date of birth will be replaced for example, with age in months*). No extra data will be recorded than the one necessary to fulfil the research objectives and will be done accordingly to the informed consent. For clinical data, we will follow a risk-based deidentification strategy, for example, this one applied to a similar database, see DOI: 10.1097/CCM.0000000000004916, figure 1. In the case of non-clinical data, no name, exact age or position will be recorded. Any other personal data provided by participants during the study that could lead to the direct identification of the participant will be minimised as much as possible.

Additionally, clinical and non-clinical **data will not be cross-linked**** (i.e. clinical data from a minor will not be matched to the end-user data from his/her caregiver or the health care professional who was assigned to the patient). Note that end-users will be given a unique identifier and therefore the same user may have several entries (e.g. the same end-user will be asked to test the system various times until usability is acceptable).

* Update data minimisation strategies during data acquisition (monitoring system) with Bering.

** Cross-linking of clinical and non-clinical data needs to be confirm by one of the research partners.

1.3 Is one of the outcomes of your project software? You can think of scripts, modules, tools, an app, a analysis pipeline etc.

- Yes

The final aim of the IMPALA project is to generate a fully functional innovative vital signs monitoring system for low resource settings. It includes:

1. an **algorithm-based software** to detect and predict critical illness real-time,
2. bedside **monitor application**, and
3. a **tablet-based application**.

1.4 How will you collect, create and/or capture your data? Briefly describe what you need to collect or access the data. Think about protocols, tools, equipment, hardware etc.

i. Non-clinical data

Non-clinical data will be collected using electronic audio/video recorders, and through notes taken by the researcher directly using a word processor or manually (and then digitalised). All data collected will then be transcribed and translated into English when necessary. Additionally, participants will record a voice "implementation diary" (done through an end-to-end encryption messaging system or through a digital recording system). All data will be ultimately stored in secure encrypted servers at TRUE. Data from Zomba Central Hospitals will directly be stored at TRUE and data from Queen

Elizabeth Central Hospital will be temporarily stored locally and then transferred to TRUE.

Questionnaires (caregiver and health staff surveys, costing data, drug inventory data) will be entered directly in KOBO; if not possible, paper questionnaires will be used and digitised by a researcher.

Usage data (recordings of the use of the user interface in terms of clicks, and interaction with the interface) will be stored as csv files.

All non-clinical data will be centralised, and stored with backup, at TRUE (Zomba Central Hospital). Data transfers will be securely done according to TRUE protocols using OneDrive to securely transfer and/or share data between sites. The Data Manager has administrative rights and will provide access and rights according to the type of user (i.e. full or partial access to documents and folders). Depending on the nature of the data, an extra security step is added using public and private keys to encrypt and decrypt the data when sharing. All institutions based in the Netherlands (AIGHD, VU, NeLL/LUMC) and the United Kingdom (Imperial College London) comply with the current EU legislation.

ii. Clinical data

Clinical data will be entered manually in an eCRF (designed using REDCap) by a research nurse during admission and during the patient's hospital stay. This data includes: patient information as recorded during admission, other relevant historical clinical data, diagnostics, treatment, concomitant medication and clinical data at discharge/death.

Additional to the standard of care, the following will be carried out:

- a blood culture and malaria slide will be performed (if not already carried out as part of the standard of care),

- venous blood samples will be taken for further laboratory testing including

-RNA and protein analysis

- to validate a POC test distinguishing viral from bacterial and parasitic disease

- a nose swab will be taken for viral analysis.

The monitoring system will automatically record vital signs data, which will be generated by the sensors attached to patients. Sensors include a ballistographic sensor, Non-Invasive Blood pressure, ECG and pulse oximeter. They will be used to detect and monitor vital signs which may include: heart rate and heart rate variability, respiratory rate, respiratory rate variability, movement, oxygen saturation, temperature and non-invasive blood pressure. The monitoring system will record all measurements during the stay of the patients in the high dependency wards (bedside monitor system and tablet*). Data includes dashboards with an overview of patients included in the study, raw data (signal traces from sensors: vital signs, patient ID, time stamp, changes in alarms), patient data for usage (all patient information: vital signs, patient ID, time stamp, changes in alarm and clinical data entries) and system usage data (alarm information -length, cause and time-, usage data and time stamp). The patient id recorded in the IMPALA monitoring system will be the same as the random patient number assigned to the clinical data to enable cross linkage.

RNA and protein will be isolated from blood samples. RNA and protein analysis will be carried out in KUHeS, located in Blantyre. Therefore, samples from Queen Elizabeth and Zomba Central Hospital will be transported, according to the laboratory protocols. Currently, a rapid test is being developed by the group of Prof. Michael Levin at Imperial College London (one of the consortium partners), this rapid test is being developed according to diagnostic standards and will be made available to IMPALA in the course of 2022. Specifics about the system and protocols used to process and generate the results of RNA and protein samples will be available to IMPALA and will be followed accordingly. RNA and protein analysis will only comprise small signatures (no high throughput data will be generated). All biomedical research activities will be recorded and performed according to Good Clinical Laboratory Practice (GCLP).

All clinical data will be necessary for training the ML algorithm. For this computationally intensive task the infrastructure and provisions of Vrije Universiteit Amsterdam (VU) will be used. For this purpose, data will be prepared according to a risk-based deidentification strategy to maintain data utility while preserving privacy. Such strategies have been successful for other similar databases (see DOI: 10.1097/CCM.0000000000004916).

At the VU an infrastructure is in place which allows for the execution of computationally demanding machine learning techniques on anonymized datasets. Often these use GPUs. Hereby, we have the following large-scale systems available:

- DAS-V (soon to be upgraded to DAS-VI), see https://www.cs.vu.nl/das5/. This supercomputer has both CPUs and GPUs available as well as large storage. This system is located on the VU premises, but can

be accessed remotely.

- SURFSara systems (Dutch National Computing Center) Lisa and Snellius are available for use by VU (guest) researchers. The systems are located in Amsterdam Watergraafsmeer and the systems are much larger. More information can be found here https://servicedesk.surfsara.nl/wiki/display/WIKI/Lisa+hardware+and+file+systems for LISA and here https://servicedesk.surfsara.nl/wiki/display/WIKI/Snellius+hardware+and+file+systems for Snellius. Access to these systems is only remote and sufficient storage is available for data.

Clinical data will be only stored temporarily outside of TRUE (i.e. in the case of the monitoring data generated in Queen Elizabeth Central Hospital**). Priority will be given to store data directly in the central database in TRUE, at Zomba Central Hospital** (i.e. all clinical data entered directly through the eCRF).

* To be confirmed if clinical data will also be entered using the tablet (additional to the interphase of the monitoring system at the bedside).

1.5 What is the size and format of your digital data? And what software do you need to collect, process and analyse these data sets?

Stage	**Specification of data set**	**Software choice**	**File format**	**Data size estimate***
Data collection				
	Clinical data sensors	IMPALA monitoring system	csv	< 2TB
	Usage data	REDCap	csv	< 1GB
	Clinical data medical records	REDCap	csv	< 1GB
	Digital voice recordings	digital voice recorder	.mp3	
	Digital video recordings	digital video recorder	.mp4	
	Observations user sessions	MS Word	.docx/pdf	
	Questionnaires	KOBO	csv	< 1GB
	Transcriptomic	rapid test Imperial College London		
	Proteomic	rapid test Imperial College London		
	Lab data	REDCap		

Raw data				>=TB
	Clinical data sensors	IMPALA monitoring system	csv	
	Usage data	IMPALA monitoring system	csv	< 1GB
	Clinical data	REDCap	csv	< 1GB
	Digital voice recordings	MS Word	.txt	
	Digital video recordings	MS Word	.txt	
	Observations user sessions	QSRNVivo	.docx/pdf	
	Questionnaires	KOBO	.cvs/.xls	
	Genetic	CMOS-MEA5000-System	.cmtr / .cmcr / .cmte]	<TB
	Proteomic	CMOS-MEA5000-System	.cmtr / .cmcr / .cmte]	<TB
	Lab data	REDCap	csv	

Processed data				10-100GB
	All clinical data (also genetic/proteomic)	Phyton	csv/sql	
	Analysed qualitative data	Atlas.ti/QSRNVivo	docx/ .pdf	
	Analysed quantitative data	Stata/R/Phyton	csv/.dta	

Results				10-100GB
	Result tables/figures	Phyton, MS Office, MS Excel	.csv/.pdf/.jpeg/.tiff	
	Processed results publication	MS Office	docx/ .pdf	

1.6 What is the estimated total size of the digital data? You can use the 'additional information' field for more details.

- >1 TB

It is anticipated that an estimated total size of **>1TB** of digital data will be generated/processed. The most data intensive activity will be the monitoring of vital signs. Approximately 4 to 6 million data points are expected to be generated for each vital sign recorded based on the targeted enrollment of 1,000 children. A previous pilot study with a previous version of IMPALA monitoring system generated approximately 2-3 GB per patient per week.

1.7 Are there any non-digital data or outputs that the project will generate?

- Yes

1.8 Please explain briefly what non-digital data or outputs you have:

In general, where digital data collection is not possible, paper will be used as a method of data collection to ensure all research data points are captured in line with the research objectives.

Non-digital outputs include:

- Notes from the implementation research: these will be transcribed and digitalised.

- Informed consent forms: will be kept safe in double lockable locations in Zomba and Queen Elizabeth Central Hospitals.

- Biological samples for standard care tests will be stored and processed according to the protocols and Good Clinical Laboratory Practice (GCLP) of the hospital where they originated (Zomba and Queen Elizabeth Hospitals). No samples or material from these samples will be stored by IMPALA.

- Biological samples (blood and swabs) for research tests (i.e. RNA, protein analysis and viral tests) will be stored at KUHeS

1.9 Will the project use existing data?

- Yes

1.10 What kind of existing data will you re-use?

- Data from academic collaborators, such as consortium partners usually with own PI
- Care data from HiX or other electronic health records (EHR)

- The project will retrospectively collect patient history using an eCRF from the patient's health passport (equivalent to the patients' health records). This will be carried out for patients in Zomba and Queen Elizabeth Hospitals.

Additionally, one of the IMPALA consortium partners, Prof. Levin's group, Imperial College London, will provide IMPALA with a scoring system to predict critical illness, which was generated in the context of the project PERFORM (https://cordis.europa.eu/project/id/668303). There are no restrictions to the use of this scoring system in the context of IMPALA.

Currently we do not foresee that PERFORM's raw data will be used directly by IMPALA. If the need to reuse this data arises (e.g. for validation of the IMPALA prediction algorithm), agreements will be established accordingly.

1.11 Is there an agreement for the use of existing data?

- Yes

1.12 What kind of an agreement do you have for the use of existing data?

- Other (please specify)

The use of the patient's health passport will be done according to the consent given by the participant's caregiver/guardian and in line with the conditions established in the protocol.

Regarding the PERFORM scoring system, no agreement is needed.

2. Data documentation

2.1 How will files and folders be named and structured?

Proposed naming convention:
‘[type]_[researcher]_[WP]_[location]_[activity]_[version]_[date]’
For each new version a new document is created.
[type] can be: raw/clean/analysis/SOP/documentation.
[researcher] initials of the creator.
[WP] work package number.
[location] Z=Zomba Central Hospital or E=Elizabeth Queen Central Hospital.
[activity] e.g. monitoring, interviews, etc. For files containing end-user data, e.g. from interviews, focus groups and research observations will include in the [research activity] field: [responder type_interview number]
[Version #] e.g. number given to the document. Incremental as there are more versions created/edited.
[date] corresponds to the date of creation in the format yyyymmdd.
Raw data files will be locked by selecting the ‘read-only’ option in the file preferences. The suffix ‘_locked’ will be added to the file name and files will be stored in the ‘_locked’ folder. Obsolete files will be moved to an ‘Archive’ subfolder of the folder the document was originally in.
Proposed folder structure:
[1]_[Project name] contains:
- [1]_[Projectname]_raw
- [2]_[Projectname]_cleaning
- [3]_[Projectname]_locked
- [4]_[Projectname]_analysis
- [5]_[Projectname]_SOP
- [6]_[Projectname]_documentation

2.2 How will versions and changes be handled?

Version numbers are incremented for each major change.
Minor changes are indicated by adding a/b/c.
GitLab will be used as a distributed version control system for version control of the codes.

2.3 Business metadata: What metadata (standard) will be used to describe the data set? Please use the 'additional information' field to briefly explain this.

- Generic metadata standard (e.g. Dublin Core)

Generic metadata standard: Dublin Core

2.4 Please describe briefly how you will create the business metadata.

We will use the Dublin Core Standard Generator:
https://nsteffel.github.io/dublin_core_generator/index.html
The consortium will consider repositories and comply with the repository requirements for business metadata.

2.5 Technical metadata: What metadata (standard) will be used to describe and/or standardize data and variables?
Please use the 'additional information' field to briefly explain this.

- Specialised metadata standard

Specialised metadata standard: SNOMED
Specialised metadata standard Genetic data: EGA metadata standards
Specialised metadata standard Proteomic data: EGA metadata standards

2.6 Please describe briefly how you will create the technical metadata.

The data dictionary will be exported directly from REDCap. Using an xls form standard to Redcap, all the variables, question labels and answer codes will be programmed using REDCap and then a data dictionary will be exported as a csv/pdf.
The AUMC dictionary will be used as a guide for the clinical database, see https://github.com/AmsterdamUMC/AmsterdamUMCdb

2.7 What supporting information and/or documentation will you create to enhance understanding of the data? Please describe briefly what is needed for peers to understand, work and/or reproduce the data.

The study protocol will be stored with all pertinent documentation (which includes original questionnaires, protocols for data collection and analysis, monitoring system specifications), after approval by COMREC (the Medical Ethical Commission in Malawi). Any changes to questionnaires and protocols will be stored as new versions and made available at the end of the study in pdf format.

A data dictionary (code book) will be available for the questionnaires and clinical data. It will be added after export of the data from REDCap for the questionnaires, and upon receipt of the clinical data. Any categorical data will be encoded with labels to give meaning to any numeric figures captured during data collection.

All syntaxes used in data cleaning and analysis (including annotation describing the goal of processing steps) will be stored to facilitate replication.

Lab journal entries from the biomedical research activities and the corresponding research protocols will be exported as pdf.

We will also include the necessary software and tools needed for reuse and state whether embargoes, licenses, commercial objectives or other conditions (like stated in informed consent agreements) have been imposed on the reuse of data.

For RNA and proteomics data, we will also write a description according to EGA standards.

A readme.txt with a list of all available files and a description of their contents will be created at the end of the project, before archiving the data.

For the implementation research data, all interview transcripts and observation notes saved as word files (and in the NVivo database) as well as the questionnaire data will include descriptions of the context in which the data were collected. This includes information on location, the data collector, sampling methodology, recruitment of respondents, time indicators and individual identifiers.)

2.8 Please tick the box to confirm that you are aware of and adhere to the applicable rules and codes of conduct for your study or project:

- **General**
  - **VSNU Code of Conduct for Research Integrity**
  - **LUMC data management guidelines**

    - **LUMC privacy policy**
- **Human research:**
    - **General Data Protection Regulation (GDPR; in Dutch: AVG)**
    - **Medical Treatment Contracts Act (In Dutch: WGBO)**
    - **Medical Research Involving Human Subjects Act (In Dutch: WMO)**
    - **Quality Assurance for Research involving Human Subjects**
    - **Code of Conduct for Medical Research (e.g. GCP)**
    - **Code of Conduct Responsible Use of Human Tissue**
- **Non-human research:**
    - **Experiments on Animals Act**

Please add an explanation when needed in the 'additional information' field.

- I'm aware of and adhere to the rules and codes of conduct that are applicable for my study.

I'm aware of and adhere to the rules and codes of conduct that are applicable for my study.

Research with human subjects will (i.e. clinical observational study and the pilot study) fully comply with the highest international standards (Good Clinical Practice and the Clinical trials directive: 2001/20/EC) and follow Malawian national guidelines (Section 18 & 48 of the S&T Act No. 16 of 2003).

Regarding research with medical devices, IMPALA will comply with international standards (ISO 14155:2011 Clinical investigation of medical devices for human subjects, and Medical Device Regulation (EU) 2017/745).

Sensitive data will follow the standards established by the General Data Protection Regulation (GDPR) (EU) 2016/679).

The IMPALA consortium will also follow the principles and rights as established in the EU Charter of Fundamental Rights, the Helsinki Declaration and the UNESCO Universal Declaration on the human genome and human rights.

2.9 Indicate which permits apply to your study and add explanation when needed in the 'additional information' field:

- Report the collection of (in)directly identifiable (research) data to the Data Protection Officer
- Approval by ethical committee for human research (METC/CCMO)

Additionally:

• Malawi bureau of standards (MBS) or Malawi pharmacy medicines and poisons board (PMPB).

3. Data storage and security

3.1 Where will you store the different parts of your digital data? When ticking the option 'other', please use the 'additional information' field to briefly explain this.

- Other (please specify)

Data will be generated from two sites: Zomba Central Hospital and Queen Elizabeth Central Hospital. Both locations will have access to a secure server to store data housed at the Training and Research Unit of Excellence (TRUE). The study team will be directed as much as possible to store all the research data in the TRUE central database (i.e. data recorded in the eCRF). In other cases, and where it is not immediately possible to store the data in the TRUE central server, data will be first stored locally and will be transferred to TRUE via cloud transfer (i.e. monitoring data and non-clinical data from implementation research). Access to consortium researchers will be arranged specifically according to their roles.

In the case of training data for the predictive algorithm, de-identified data will be transferred to the Vrije Universiteit Amsterdam after following a risk-based de-identification strategy. This strategy will preserve privacy while maintaining data utility while preserving privacy. Such strategies have been successful for other similar databases (see DOI: 10.1097/CCM.0000000000004916).

3.2 Please describe how safe storage is guaranteed for each part of your data during collection, storage and sharing of data: storage location, backup procedures, frequency and who is responsible.

The TRUE database will be backuped every day at 15:00 GMT+2. The backup format for each daily backup will be yyyymmdd_redcap.sql. In addition to the local backup on the server, a copy of the backup will be transferred to an external hard drive housed at the TRUE main office and an additional copy will be uploaded to a secure encrypted cloud drive. The data manager at TRUE is responsible for ensuring that backups are conducted as per their ICT policy and inline with this DMP.

Data from the monitoring system will be stored in the central database at TRUE. Where it is not immediately possible to store the data in the TRUE central server due to unforeseen circumstances, data will be stored locally and transferred at the soonest time via the cloud to TRUE.

For the analysis of **qualitative data**, QSR NVivo automatically creates backup 'recovery' files automatically. These files will be backed up daily. A recovery copy of all transcripts and fieldnotes will also be saved as MS word documents separate from the QSR NVivo database, which will not be used in the analysis process (in case of file corruption during coding in QSR NVivo).

Biological samples will be securely stored according to the standard procedures of KUHeS. Samples will remain traceable (e.g. in case that a participant withdraws). Samples will be stored at the KUHes archive -80 ℃ refrigerators and samples ID will be traceable through the KUHes archiving laboratory information management systems.

3.3 For non-digital data, please specify briefly where you will store these non-digital data and describe who is responsible for handling and storage of these outputs.

Written informed consent will be stored in a double lockable safe location with cabinets design specifically for this type of research documentation), and in the institution where they were obtained (i.e. Queen Elizabeth and Zomba Central hospitals). The responsible for clinical research (*Dr Jenala Njirammadzi, KUHeS) will be responsible for ensuring the right storage of informed consent forms.

Non-digital data from implementation research (research notes, observations, voice recordings and videos) will be digitalised and stored at the end of data analysis and dissemination to allow for additional validation and quality control. All non-digital data/participant records will be archived at the TRUE archive upon the completion of the study. Prof. Wendy Janssens (AIGHD), the leader of social sciences and implementation research will oversee the correct handling and storage of this non-digital data.

Biological samples (blood, swaps and isolated RNA and protein) will be stored at KUHeS in -80 freezers. Storage will follow the SOPs of KUHeS.

*Leaders of Clinical Research and Social Sciences research will be confirmed in the next Management Team Meeting in January 2022.

3.4 How will access to data be managed during the project?

Please specify for each storage device the tools and procedures that you use to ensure that only authorized persons have access to data. Outline roles and responsibilities for all activities during your project, e.g. data capture, metadata production, data quality, storage and backup. For collaborative projects you should explain the coordination of data management responsibilities across partners.

Clinical data will be entered in eCRFs by healthcare professionals. The monitoring system will comprise the sensors that will be connected to the bedside monitor (where they can be visualised) and which will automatically store all data by wireless connection to TRUE or a local server where immediate transfer is not possible. Data from several monitors will be connected wireless to tablets where an overview of all patients using the monitoring system will be displayed. In addition clinical information may be collected directly through the tablet which will also be transferred wirelessly to the servers.

Laboratory results of standard care will be entered in the eCRF by a research nurse.

Protein and RNA data will be generated at KUHeS and entered in the eCRF by a researcher.

Biological samples taken solely for research purposes (blood samples and swabs) will be stored according to KUHeS laboratory SOP and will only be accessible to the project researchers under the supervision of Dr. Myrsini Kaforou (Imperial College London).

CRFs and the database will be designed by the PhD researcher MSc. William Nkhono, and will have the support of a data clerk based at TRUE. MSc. William Nkhono will act as PhD researcher and data manager of the project and will receive support of Prof. Kamija Phiri (TRUE) and Prof. Mark Hoogendoorn (AIGHD) and PhD María Villalobos (NeLL, LUMC), and the Advanced Data Management Group (LUMC).

The PhD student and data clerk will oversee data capture, validation, transfer, analysis and storage; across the whole project. All data will be centralised at TRUE. Access will be granted to each consortium researcher according to their role by the project data manager. Support will be received from the logistical “hub” leaders*:

- for biological samples Dr. Myrsini Kaforou (Imperial College London),
- for dynamic vital signs and IMPALA monitoring system MSc. Bart Bierling (GOAL 3),
- for clinical data Dr. Jenala Njirammadzi (KUHeS),
- for social sciences and implementation research Wendy Janssens (AIGHD).

*Hub leaders will be confirmed during the January 2022 Management Team Meeting.

3.5 Do you have a plan or SOP for quality control of your data? Please explain briefly in the 'comment area'.

- Yes

Quality checks (validations) will be implemented in REDCap to improve data entry quality. REDCap allows for basic validation steps, including real-time edits checks and field-logic checks such as valid range and outliers/expected values. All data collection tools will be extensively tested to ensure integrity.

During data collection, trained personnel will act as a quality control measure verifying their work when collecting the data and before saving it. Before any data is submitted to the TRUE server, a coordinator/data clerk will verify the data collected (i.e. number of records, any typos, etc) and then transfer it to the server.

Additionally, at Queen Elizabeth and Zomba Central Hospitals an internal monitor will review a randomly selected 20% proportion of patients recruited that month. Patients or fields with errors will be queried with nurses and measures will be taken accordingly to correct the data entry and prevent future errors. Where necessary, further training will be provided when deemed appropriate. based on the findings of the random selection.

Data cleaning steps will be described in the analysis plan and will be performed by the researcher and supported by the data manager. Interim analyses on recruitment and follow-up done by the data manager with assistance from scientific staff will also serve to identify discrepancies and generate queries for quality control.

Transcriptomic and proteomic analyses will follow quality standards as described by the manufacturer of the quick test (Prof. Levin's group, Imperial College London), any samples that do not meet the established standards will be discarded. All laboratory activities will be recorded in electronic laboratory notebooks, which can eventually be checked by the supervisor.

Qualitative data (transcripts, translations and coding) will be always reviewed and validated by a second researcher, as well as a randomly selected 10% of the questionnaires (caregiver surveys).

At a higher level, data generated by all IMPALA teams will be reviewed during general meetings with

the PI Dr. Job Calis. Summarised statistics will be validated and in case of irregularities measures will be taken. Monitoring activities (regarding Good Clinical Practice and data monitoring) will also ensure the quality of data. The data management team will monthly perform a site monitoring, followed by a study close-out visit. Reports will be discussed, and improvement measures will be taken accordingly. Before the definitive data lock, at least two interim locks of the central database will be carried out for interim analysis (when achieving 50% and 75% of patients enrolled), or at request of the monitoring team.

3.6 Do you expect costs for storage and data management during the study or project?

- Yes

3.7 Please describe briefly how these costs will be covered. If you have budgeted for this in your grant application, please specify.

Grant application has budgeted 22.500euros for data analysis and storage, including indirect costs.

4. Data access, sharing and reuse

4.1 Are there any restrictions placed on sharing/reuse of some/all of your data due to one or more of the following options? When you tick the box 'other', please specify this in the 'additional information' field. You can also use this field to give more information.

- Other (please specify)
- Consortium agreement
- Intellectual Property (IP) e.g. patent

Clinical data generated during this study is considered highly sensitive data. Storage of the personal data will be confined to secured databases at TRUE and personal data will be de-identified prior to sharing with consortium members, ensuring the protection of privacy of participants. Sharing data within the consortium activities will be carried out according to the consortium agreement, which will consider the conditions established by the GDPR. DTAs will be signed when necessary.

The monitoring system is being developed by a social enterprise (GOAL 3), which is part of the consortium. It will be established if an embargo period for sharing data is necessary and if the monitoring system, software and algorithms will be subjected to IP rights.

Informed consent will include the choice to consent to sharing data beyond the project with restricted access, for secondary purposes, and only when certain conditions are met. We do not plan to restrict use of data depending on the nature of the requester (public or private), or objectives (for-profit or non-profit). Access will be granted according to the conditions established in the consortium agreement and by a data access committee.

4.2 Will you share your data open access or with restricted access?

- Restricted access

4.4 Sharing data with 'restricted access': please explain if this is done to publish or seek for patents, or because your data contains privacy-sensitive information. And how will you share data under 'restricted access'?

Restricted access is chosen mainly because of the privacy-sensitive information (clinical data, minors, in a low income country setting). There are some considerations about IP that have been defined by the consortium in the beginning of the project.

4.6 Is there an embargo period before sharing your data?

- Yes

4.7 Why and for how long do you have an embargo period?

Conditions for embargo periods have been established in the consortium agreement. Before publication, an embargo period may be applicable in order to protect the database, and in line with achieving the goals of the consortium. Metadata however will be made available immediately.

4.8 Is your informed consent form according to the LUMC-based CCMO model form? If no, please explain in the 'additional information' field why you don't use this standard.

- No

The model form is the one designed by COMREC, the College of Medicine Research Ethics Committee (Malawian institution), the ethics committee that oversees the two medical institutions involved in this study.

4.9 How do you ensure that participants, who have withdrawn their informed consent, are removed from the data and thus are not available for reuse? Do you have a procedure in place for this?

Biological samples (i.e. blood, nasal swabs, RNA and protein samples) will be traceable and it will be possible to destroy them upon withdrawal of consent. All samples will be centralised at KUHeS and the process of destruction of samples will be coordinated and overseen by the leader of biomedical sciences research: Dr. Myrsini Kaforou (Imperial College London)*.

The layout of the database will foresee an erasure protocol, as an automated data validation step. With all appropriate supporting documentation of the participant's claim to withdraw their informed consent from the study, a script will be written to drop/delete any and all data related to the participant stored in the database. In addition, the participant chart will be labeled appropriately to visually indicate to any research personnel of the participant withdrawal. The file will be moved to a separate location where participants' records who have withdrawn are stored away from the active participant records. Accordingly, no more data will be generated after withdrawal of consent.

Note that regarding the machine learning algorithms however, deleting the data will only ensure that the data will not be used in the future. The models that are built will however be on such a level that it will not be possible to distill the original data of the participant from the models.

* To be confirmed during the Management Team Meeting in January 2022.

4.10 Does your agreement or funder requirements include information about intentions for

sharing, retention of data, steps taken to protect participants privacy, confidentiality and ownership of data and intellectual property rights?

The EDCTP2 programme is funded under the Horizon 2020 programme (H2020) and is committed to open access. Open access refers to the practice of providing online access to scientific information that is free of charge to the end-user and reusable. This encompasses:

- Peer-reviewed scientific research articles (published in scholarly journals)
- Research data (data underlying publications, curated data and/or raw data).

In the case of research data: metadata will be made openly available, de-identified data will be shared under restricted access and raw data is considered highly sensitive data and in principle it will not be shared with third parties to ensure data privacy and protect participants of the study. This will be done according to the policy on registering and reporting clinical studies of EDCTP.

For further details refer to:
http://www.edctp.org/web/app/uploads/2018/07/EDCTP2_policy_on_registering_and_reporting_clinical_studies-1.pdf

4.11 Who is responsible for your data and has authority to grant (additional) access to your data after finishing the study or project (e.g. for the long term)?

- Data Access Committee

During the duration of IMPALA project, IMPALA's Core Management Team will decide whether to grant additional access to third parties, an equivalent committee will be put in place after the completion of the project.

5. Data preservation and archiving

5.1 Which parts of your data will you select for long-term archiving? Please motivate why you would not archive (parts of) your data.

Archiving data has two different goals, reproducibility/transparency and future data sharing.

After the completion of the study all essential documents will be preferably stored in digital format at TRUE. Essential documents include the protocols, Medical Ethical Commission documentation, product information (regarding the monitoring system, i.e. sensors and hardware) and all the patient and participant data.

Documents with hand signatures (e.g. the informed consent and the protocol's signature page) will be scanned and digitally stored, the original counterparts will be kept in safe locations (will not be destroyed).

5.2 How long must your data be preserved? Please explain briefly in the 'additional information' field.

- Clinical research WMO: ≥ 20 years

In line with the Netherlands Code of Conduct for Research Integrity and Malawian best practices, raw and processed data will be stored for a period of at least 10 years.

5.3 Are there any requirements regarding the disposal of data?

- Yes

5.4 What are the requirements regarding the disposal of data? Describe how you will dispose of the data: how you will get approval, what people and/or tools you need, etc.

Biological samples will be disposed of following KUHeS waste protocols.
Raw data will be disposed of according to TRUE standard procedures.
Informed consent forms will be disposed of only after no biological material is available. The documents will be disposed of according to Queen Elizabeth and Zomba Central Hospitals procedures.

5.5 How will you ensure data and/or metadata findability and availability for the long term?

Briefly explain your choices in the 'additional information' field, in which you specify how you ensure long term data availability. If you don't deposit in an established repository, you should explain what resources and systems are in place to enable data to be curated effectively beyond the lifetime of the project.

- I will publish my metadata online
- Archive software and scripts on a specialised platform
- Archive (parts of) data in a field-specific database/archive/repository

All study metadata will be published and shared in an online repository.
Archive software and scripts, when not subjected to IP, will be shared on a specialised platform.
Archive (parts of) data in a field-specific database/archive/repository which can be publicly accessed for data reusability.

5.6 Does the chosen publication format (database, archive, repository, catalogue, platform, website etc.) add one or more persistent identifiers to your (meta)data?

Please specify the type of persistent identifier(s) in the 'additional information' field.

- Yes

Although no repository has been chosen yet, assigning persistent identifiers to metadata is common practice. We will ensure the chosen repository uses persistent identifiers.

5.7 What will you do to prepare your data for archiving? Describe how you intend to meet LUMC, publisher or database/archive/repository requirements.

All files will be converted into preferred formats according to the repository or database requirements. Only metadata will be made available through repositories. The consortium will follow FAIR principles and the principle as open “as possible and as close as necessary”.

5.8 Will there be extra costs for this preparation? If you have budget for this in your grant proposal, please specify.

- Yes

5.10 Do you have costs associated with long-term storage of your data?

- Yes

5.11 How will these costs for long-term storage be covered?

The consortium will discuss the conditions for long-term storage. TRUE is prepared to store all research related data for 10 years.
Additionally, the consortium aims to, after the completion of this study, set up the conditions to build a restricted access database. This database will only include de-identified data and is meant to facilitate research and development of healthcare solutions for paediatric settings in LRS. Funding beyond IMPALA will be discussed at least one year before the completion of the study to ensure the continuity of the database and its access to the research community.
Under the IMPALA project, samples will only be transferred within Malawi. By the end of the project, the remnant biological samples (blood samples, swabs, and isolated protein and RNA samples) will be aliquoted and may be shared between consortium partners. It is a priority that a fraction of the aliquots remain in the hands of African partners to support future research on-site, IMPALA will support capacity building if necessary. European partners that may receive samples will be established in the DMA and will have to demonstrate appropriate quality certification standards. Cost for shipment within Malawi is included in the IMPALA budget, for transfer beyond this project the receiving institutions will cover storage expenses. The informed consent will include the relevant information in this regard, and will inform participants' representatives about their rights, including the right to access information, trace and/or withdraw samples/data.

6. Additional information

6.1 Here you can put any additional information that you were not able to list above.

Question not answered.

7. Review request

7.1 Please tick the appropriate box.

- review: funder requirement

Feedback

Feedback and suggestions are more than welcome!

Question not answered.