

School of Computer Science

Data Management

in Fulfilment of

SPEC9997

Maksymilian Drzezdzon

C15311966

Degree: TU060/1

Module Coordinator: Emma Murphy

Declaration of Ownership: I declare that the attached work is entirely my own and that all sources have been acknowledged: 🗹  
**Date: 2021/03/24**

Data Management Plan &

Data Protection Impact Assessment



***Maksymilian Drzezdzon***

*C15311966*

This Project submitted in partial fulfilment of requirements of Technological University Dublin for the degree of

M.Sc. in Computer Science

May 2022

Table of Contents

[Data Management Plan 1](#_Toc69654496)

[Data Flow 1](#_Toc69654497)

[Handling Data Quality 2](#_Toc69654498)

[Accuracy 2](#_Toc69654499)

[Completeness 2](#_Toc69654500)

[Consistency 2](#_Toc69654501)

[Data Integrity 2](#_Toc69654502)

[Reasonability 3](#_Toc69654503)

[Timeliness 3](#_Toc69654504)

[Uniqueness 3](#_Toc69654505)

[Validity 3](#_Toc69654506)

[Data Bias 3](#_Toc69654507)

[Data Privacy and Security 4](#_Toc69654508)

[Data Protection Impact Assessment 5](#_Toc69654509)

[Processing Personal Data 5](#_Toc69654510)

[Data Stakeholders 5](#_Toc69654511)

[Data Safeguards 5](#_Toc69654512)

[Data Security 5](#_Toc69654513)

[Data Consent 6](#_Toc69654514)

[Ethical Risks 6](#_Toc69654515)

[References 7](#_Toc69654516)

# Data Management Plan

## Data Flow

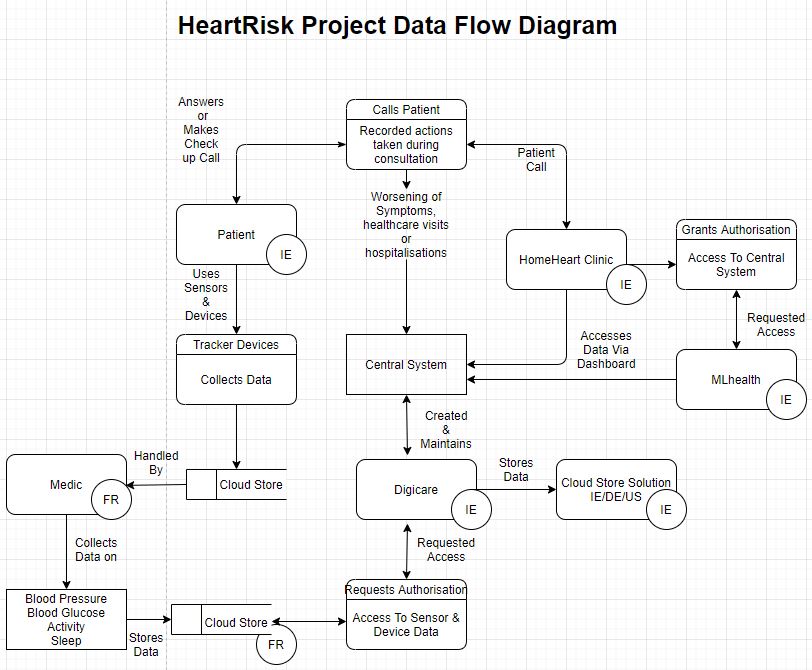


Figure 1: Data Flow Diagram

Digicare has data centers in Ireland, Germany and the United States. Their data center based in Ireland should be utilized for this project, keeping data within the EU and the country in which the project is being carried out.

The initial data collection is handled by Medic via their medical devices. This data will need to be access by Digicare for further processing. With such a broad range of data stakeholders, the proposed solution would be to use a private network, which HomeHeart clinic and MLhealth would need a VPN to access, it would also need to be configured by Digicare. Further details on this will be described in further sections in data security and privacy.

The data types for the above diagram would consist of text for names, addresses among anything else that is supplied, I’d need to request a list column to provide a more extensive list. Anything on medical records would also be foreign to me, a short brief on that would also be required for the purpose of the evaluation.

The sensor data gathered by Medic devices would be numeric in nature. All of this would be rendered on the Digicare systems dashboard accessed by HomeHeart Clinic and MLhealth.

## Handling Data Quality

It’s important to define what high and low-quality data are. High quality data is data that meets the expectations and needs of a data consumer, if it is not suitable for its intended purpose, its low quality. (International, DAMA-DMBOK: Data Management Body of Knowledge: 2nd Edition, 2017)

Data quality can be assessed using data quality dimensions outlined by DAMA International. (International, DAMA-DMBOK: Data Management Body of Knowledge: 2nd Edition, 2017)

Data that is being recorded is:

blood pressure, activity trackers (sleep, activity), weight, health condition via phone consultations. Twelve months of medical history data. Age, gender, heart condition among other health conditions such as angina, partial or indefinite sight impairment, mild or severe hearing loss or disabilities such as mild or severe intellectual disability, physical disability.

### Accuracy

Data mentioned above seems to reflect real life circumstances well, one criticism would be that sensors and IoT devices are prone to relaying data inconsistently at times, leaving gaps.

### Completeness

Similar to the statement above, IoT devices would need to be monitored for failures with replacements ready or already supplied and configured.

### Consistency

Again, IoT devices would be the main vector of failure, assuming that previous records and information is complete. There is some tolerance for missing data from said IoT devices which can be filled in with averages or otherwise.

### Data Integrity

Having such a vast collection of various health records leaves researchers replete with ways to validate model results with past medical history and current trends presented on dashboards.

### Reasonability

Relaying back to the above statement, there is enough data that it’s possible to cross check results and trends in real time. There is a concern with the 2:1 gender split among other details, more on that later.

### Timeliness

This is a difficult criteria to address, generally, data should be rendered almost in real time in respect to potential gaps as described above, its validity and longevity shouldn’t be affected. However, participants wellbeing needs to be taken into consideration.

### Uniqueness

There will be potential duplicates from sensors, someone can exhibit similar patterns of movement or heart rate, blood pressure etc. This should be expected.

### Validity

See above. Most concerns have been highlighted with ways to cross check information.

## Data Bias

There is an uneven split of male to females a ratio of 2:1. This can be either fixed with adding an additional weight to balance out the data, build separate models for classifying heart risks in females or trying to reduce the number of participants in the male group meanwhile potentially removing patients that pose an ethical risk, such as those with cognitive impairment, more on that later.

Another method would be asking which columns are needed, this could lead to further homogenizing of the dataset assuming that heart disease symptoms are alike for both groups, I’m not a clinician hence I’m not able to speak to that.

I know in mental health both genders do exhibit different severities of behaviors such as hyperactivity or inattentiveness in ADHD etc., with a clinician’s review, if heart symptoms are alike in both groups, then it can be homogenized through anonymization, something that would be required regardless.

Interpretability of models in machine learning applications such as this is key, hence linear and generalized regression models could be used among other white box models that are interpretable which will help in understanding of results.

There is a parameter called weight which will only be required for ‘Heart failure patients’ I’m not particularly sure why that’s needed and how or if that will skew the analysis, since this data wouldn’t exist for other patients. That question would need to be answered along with potentially collecting similar data from other patients as a control for this analysis.

Its not particularly clear whether table 1 in the report envelops 1000 patients, do these conditions overlap for 1 patient? Or are the individual. This makes it difficult to judge the spread of said conditions and how that might contribute to the analysis.

## Data Privacy and Security

The proposed security solution can be handled via a dedicated private network maintained by Digicare. In order to connect to this private network users must use a VPN that could need to be configured by Digicare. In addition to that, each person logging in must have an account used to log into the VPN and services to view the dashboards etc but also an authentication key also known as a hardware key. This key would be synced with a simple internal app to match key codes when it’s connected to a computer and pressed for an input.

Regarding personal IoT devices, a system should be put in place to help individuals from becoming victims of cybercrime, preferably one that doesn’t require any input on their part.

A possible suggestion could be loading up configuration locks on devices along with securing their home networks, if possible educating participants with a quick digestible video and/or transcript depending on preference, overviewing basic security etiquette.

As to data privacy, only authorized staff should be able to access required databases. Once accessed personal health records should already have had data extracted and joined with associated patient before being anonymized. Meaning, any personal info should be omitted, thus being unavailable to everyone. The only staff members with access to personal health records should be nurses that conduct check ups and offer consultations, encased and accessed through the system outline above.

Data should not be held on the same server; in case one is compromised only a fraction of said data should be lost with it. An individual may receive verbal communication of their biometric readings from a nurse that has access to their dashboard via phone consultations.

# Data Protection Impact Assessment

## Processing Personal Data

When processing personal data in Europe or from a European country a company must abide by the GDPR. Meaning, only processing what is needed and keeping it only as long as its needed. Data should be anonymized or pseudonymized, replacing personal identifiers entirely or with artificial ones. (Irwin, 2020) As mentioned in the project description, Digicare stores its data on secure servers which for the sake of this project will be located in Ireland, it is recommended that this data is encrypted and only accessible via the system outlined above, with a limited number of members holding accounts that can directly access it.

Any identifiers that can be used alone or in conjunction to distinguish an individual such as name, id number, location, IP address need to be hidden or removed. In this project biometric/health data can also be used to identify someone.

Consent must be requested from each individual to process their personal data, outlining that their medical data is going to be used for what purposes. Participants that have a cognitive disability require a guardian to provide consent, more on that later.

## Data Stakeholders

2. If the project involves multiple organizations, identify the data controller(s) and

processor(s)? (Justify your answer)

[10 marks] – 250 words

Patient – Data Stakeholder

The patients own their data and have provided consent to provide it over a two-year period for the sake of the HeartRisk project.

Medic – Processor

Medic is a French company that is operating on the behalf of MLhealth’s team meanwhile providing data to Digicare, as they have the needed infrastructure for process and thus control said data.

Digicare – Processor/~~Controller~~

Digicare is a data controller in regards to ensuring security and providing infrastructure

MLhealth – Processor

HomeHeart Clinic – Processor

HomeHeart provides consultation services for patients, they can access data to provide advice on steps a patient should consider taking to ensure their wellbeing. They don’t augment or change data in any way.

## Data Safeguards

3. How will you apply safeguards to ensure the processing remains lawful e.g.

Pseudonymization, anonymization?

[10 marks] – 250 words

## Data Security

4. How will personal data be secured throughout its entire lifecycle?

[10 marks] – 250 words

## Data Consent

5. If relying on consent to process personal data, how will this be collected and what is

the impact if consent is withheld or withdrawn?

[5 marks] – 125 words

## Ethical Risks

6. What are the critical ethical risks for this project and how can you mitigate for them.

[10 marks] – 250 words

# References

International, D. (2017). DAMA-DMBOK: Data Management Body of Knowledge: 2nd Edition. In D. International, *DAMA-DMBOK: Data Management Body of Knowledge: 2nd Edition* (pp. 449-494). Technics Publications.

International, D. (2017). DAMA-DMBOK: Data Management Body of Knowledge: 2nd Edition. In D. International, *DAMA-DMBOK: Data Management Body of Knowledge: 2nd Edition* (pp. 123-166). Technics Publications.