

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](#_Toc69739617)

[Data Flow 1](#_Toc69739618)

[Handling Data Quality 2](#_Toc69739619)

[Accuracy 2](#_Toc69739620)

[Completeness 2](#_Toc69739621)

[Consistency 2](#_Toc69739622)

[Data Integrity 2](#_Toc69739623)

[Reasonability 2](#_Toc69739624)

[Timeliness 3](#_Toc69739625)

[Uniqueness 3](#_Toc69739626)

[Validity 3](#_Toc69739627)

[Data Bias 3](#_Toc69739628)

[Data Privacy and Security 4](#_Toc69739629)

[Data Protection Impact Assessment 5](#_Toc69739630)

[Processing Personal Data 5](#_Toc69739631)

[Data Stakeholders 6](#_Toc69739632)

[Data Safeguards 7](#_Toc69739633)

[Data Security 8](#_Toc69739634)

[Data Consent 9](#_Toc69739635)

[Ethical Risks 9](#_Toc69739636)

[References 10](#_Toc69739637)

# Data Management Plan

## Data Flow

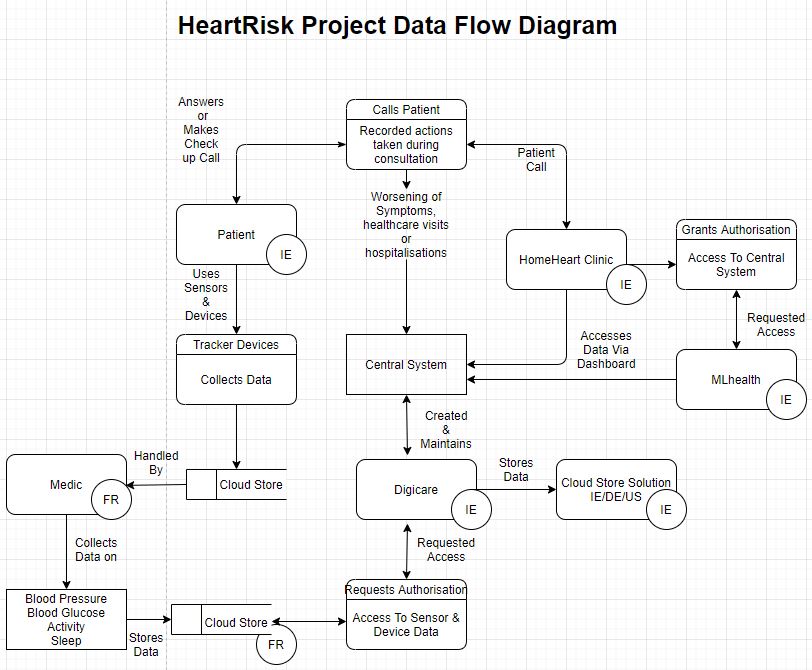


Figure 1: Data Flow Diagram

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. (Bagdasarian, 2021)

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. (Bagdasarian, 2021)

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.

When too many requests are made on an account said account should become locked with appropriate members of the security team and account owner being alerted. (Bagdasarian, 2021)

# 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

**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/~~Joint-Controller~~**

Digicare is a data controller in regards to ensuring security and providing infrastructure however they act and work under the MLhealth group making Digicare a Processor and MLhealth the controller.

**MLhealth – Controller**

MLhealth are the controllers in this project, they use and make decisions on how data is going to be used as they are the ones building the machine learning project and using said data for research purposes. Any other entity involved is operating under MLhealth’s instruction.

**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

Personal Health records should only be accessible to nurses with the consent of each patient when providing consultations. Much of this has been outlined above in the security and data privacy section.

To access any dashboard each nurse needs to connect to the Digicare network with a VPN, login with their account and use a hardware key to establish the initial connection, then said dashboard can be viewed. The dashboards session should time out after 25 - 30 minutes logging the user out requiring a fresh code/token to be provisioned from the hardware key.

Likewise for research staff, they will need to follow the same authentication process. Only research staff should have direct access to servers that hold data for the machine learning project. (Bagdasarian, 2021)

Research staff can use Digicares cloud systems to run their code and preform any transformations needed without having to download any data. Each participating staff member should also undergo appropriate HIPAA and data protection/handling training.

Each staff members laptops that have access to patient data should be encrypted locking it if its stolen or misplaced. (Bagdasarian, 2021)

## Data Security

**Data Creation**

Once patients receive their IoT devices, data collected is stored on Medics cloud solution. These devices should work out of the box and require no additional input to prevent any tampering with said devices.

**Storage**

No one should be accessing this data other than authorized data engineers from Digicare when they pull it to their infrastructure. It will be stored on site with database hosted on secure servers located in Ireland. (Kumar, 2020)

**Usage**

Data Availability, dashboards rendered for HomeHeart clinic will be updated as close to real time as possible with appropriate data pipelines being put in place to automate this process assuring its security and autonomy.

**Archival**

As mentioned previously, its recommended that data operations happen on Digicares cloud solution with data and models being modified without the need to be downloaded. Appropriate data resiliency best practices should be reviewed so that multiple tables and/or data stores don’t grow over time, meanwhile removing or discarding any data that isn’t needed. (Kumar, 2020)

**Disposal**

In a case that a user wishes not participate in the study, their data should be purged, its important that that number of copies of said data are kept to a minimum in the data archival stage to speed up relaying processes.

## Data Consent

Consent forms can preferably be filled in person either in the MLhealth office/place of operation. There a collection of brochures or other informative documents could be given that each participant can read, consisting of what the project is, what it’s trying to accomplish etc. (GDPR-EU, 2021)

When a patient can’t travel to said facility either a phone call can be arranged or whenever possible someone could be dispatched with the patients consent to fill out forms at their home.

Research staff would then fill out appropriate forms with participants answering any questions assuming they have the capacity or knowledge to do so, otherwise a follow up email or phone consultation could be used to fill out knowledge gaps.

In the advent of a patient not wishing to participate anymore a phone call should be made, in that scenario data will need to be removed within 30 days to comply with data protection law. Section 12.3 (Wolford, 2021)

## Ethical Risks

Deontological ethical frameworks, place emphasis on the characteristic of the action itself and not on the product of said action. (Britannica, 2021)

One ethical risk that instantly jumps out is that some participants seem to have some level of cognitive impairment, for the sake of their wellbeing even if permission is granted it would be recommended that they are omitted from this project. Reason being is that tracking someone who’s incapable of making these decisions for themselves seems unethical and doesn’t contribute to the project. Other participants that better fill this role could be attained.

Furthermore, two participants are blind with thirty-three having a physical disability. Blind patients should be omitted for similar reasons to those with cognitive impairments. Putting vulnerable people’s private lives at the risk isn’t necessary.

Another thing worth mentioning is that patients with physical disabilities will exhibit much less movement than the rest of participants. Equipment provided to them might not be suitable and could potentially lead to injury depending on circumstances.

An argument could be made that these patients could provide data on their subgroup of elderly people that are at an even higher risk of heart disease because they probably move around much less. Having said that, data received from these patients would need to be heavily weighted in order to draw a correlation. This isn’t necessarily an ethics issue more a safety issue, appropriate adjustments would need to be made with regards to equipment provided to suit each person’s physical disability if they wish to participate so their safety is

assured.

# References

Bagdasarian, H. (2021, April 19). *Data Protection Safeguards Categories*. Retrieved from Identity Theft Awareness: https://www.identity-theft-awareness.com/data-protection-safeguards.html#:~:text=Technical%20Safeguards,intrusion%20prevention%20or%20detection%20controls.

Britannica, T. E. (2021, April 19). *Deontological ethics*. Retrieved from britannica.com: https://www.britannica.com/topic/deontological-ethics

GDPR-EU. (2021, April 19). *Everything you need to know about the “Right to be forgotten”*. Retrieved from https://gdpr.eu/: https://gdpr.eu/right-to-be-forgotten/#:~:text=In%20Article%2017%2C%20the%20GDPR,originally%20collected%20or%20processed%20it.

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.

Irwin, L. (2020, Nov 12). *The GDPR: What exactly is personal data?* Retrieved from itgovernance.eu: https://www.itgovernance.eu/blog/en/the-gdpr-what-exactly-is-personal-data

Kumar, S. (2020, July 21). *WHAT IS DATA LIFECYCLE MANAGEMENT?* Retrieved from stealthbits.com: https://stealthbits.com/blog/what-is-data-lifecycle-management/#:~:text=Data%20Lifecycle%20Management%20Definition&text=Data%20lifecycle%20stages%20encompass%20creation,%2C%20resiliency%2C%20and%20regulatory%20compliance.

Wolford, B. (2021, April 19). *Art. 12 GDPRTransparent information, communication and modalities for the exercise of the rights of the data subject*. Retrieved from https://gdpr-info.eu/: https://gdpr-info.eu/art-12-gdpr/