|  |  |
| --- | --- |
|  | |
| Data Management–Assessment 2  TU060 : Case Study: MLHealth/TriCARE – *Predict* Project.  Data Management Plan  Data Protection Impact Assessment | |
|  | |
| Ciaran Finnegan – Part Time – First Year 2021/2022  MSc in Computer Science (Data Science)  Student No : D21124026  11/5/2022 |  |
|  |  |

Table of Contents

1 Introduction 3

1.1 Purpose of Report 3

1.2 The Data Management Plan 3

1.3 The Data Protection Impact Assessment 3

2 Data Management Plan 4

2.1 Project Predict: Data Flow Lifecycle 4

2.2 Project Predict: Data Quality Issues 5

2.3 Project Predict: Data Bias and Remediation 7

2.4 Project Predict: Data Privacy and Security 9

3 Data Protection Impact Assessment 10

3.1 Project Predict: Basis for Lawful Data Processing 10

3.2 Project Predict: Data Controllers and Processors 11

3.3 Project Predict: Safeguards 14

3.4 Project Predict: Data Collection and Consent 15

3.5 Project Predict: Ethical and Privacy Risk Matrix 16

4 References 24

# Introduction

## Purpose of Report

Project Predict has achieved ethical approval in principle but part of the final funding process involves the generation of a Data Management Plan (DMP) and a Data Protection Impact Assessment (DPIA).

This document contains both the DMP and DPIA content and is initially intended for internal project sponsor review.

The DPIA (Section 3 of this document) makes a number of recommendations for Project Predict to consider in terms of appropriate and necessary data risk management processes.

## The Data Management Plan

The DMP is a formal document that explains how data should be managed through the lifecycle of the Project Predict initiative.

## The Data Protection Impact Assessment

The DPIA is a formal documentation of the risks that can arise out of the processing of patient data throughout the Project Predict lifecycle. It also includes remedies and actions to identify and minimise these risks as early as possible.

# Data Management Plan

## Project Predict: Data Flow Lifecycle

System data flow diagrams can often follow a Context Diagram/Diagram 0 format**[1]**, but for our purposes we need to understand the lifecycle of data as it moves through the ***Predict* Project**.

The diagram below borrows from the circular format used by Folio3 Dynamics**[2]** to represent the key data lifecycle phases for the project. Stakeholders and individuals are marked in ***Italic Blue*** and the associated data types that are input/output by each phase are included.

Timeline

Description automatically generated

Figure 1 – Predict Project Data Lifecycle

For the purposes of this report the ML process is considered as an independent entity. It generates a triage recommendation and will ultimately function as a proxy for assessments conducted by nurses.

## Project Predict: Data Quality Issues

Below are a brief list of potential data quality issues and remediations, with which Project Predict should have a strategy to tackle;

*Publish a formal PDCA Model*

Be upfront and declare that a continuous commitment to data quality is a central tenant for Project Predict throughout the entire four-year lifecycle. A documented Plan-Do-Act-Check (PDCA) model will inspire confident that the sensitive medical data held on patients by Project Predict will be maintained with the greatest level of possible integrity. (Taylor et al., 2013) provide excellent guidelines in how to apply this in a healthcare setting**[3]**, such as Project Predict.

*Ensure Sensor Accuracy*

DigiHealth must incorporate appropriate validation routines into the system implemented for Project Predict so that faulty Medic sensors cannot introduce erroneous data, with values outside of possible medical ranges. The potential challenge with such Medical Internet of Things (IoT) is one of the issues discussed in the 2019 paper by Krishnan and Shasidhar**[4]**.

Similarly, sensor data must always be complete. Missing health metrics should also generate a system alert for correction.

*Realtime Update of Patient Dashboards for Nurses/ML Process*

Timeliness is a critical issue in the triage alerting process for Project Predict. Sensor data should be transmitted in real time, and not through a scheduled batch update. Patient dashboards that may need urgent attention should render as quickly as possible, with supplementary alerting to nurses/ML systems if immediate diagnosis is recommended. Senor data must also be timestamped to confirm that it is the most recent copy of data, and also allow a future trend analysis.

*Data Store Integrity*

The data model within the TriCARE datastore will store daily sensor data, triage recommendations and outcomes, and then, at a later stage, the Year Three Trial survey information. All of this information, which is from diverse sources, must seamlessly tie into a single patient data entity in the TriCARE systems, as implemented by DigiHealth.

*Data Audits*

The project needs to avoid an ad-hoc approach to checking data quality, and reliance on TriCARE employees conducting occasional reviews to capture data inconsistencies. A formal ‘data audit’ should take place each month by nominated TriCARE team members, which rewards the discovery of data issues and/or suggestions around data quality management.

*Data Storage*

Lastly, to avoid data synchronisation issues it is strongly encouraged that there is ‘one version of the truth’. TriCARE should manage a single datastore, through the DigiHealth cloud infrastructure, and allow appropriate access to MLHealth researchers. There is research on current options to allow data replication across healthcare sites**[5]** but in this is likely beyond the scope of Project Predict and introduces a possibility of unnecessary data duplication and data consistency errors.

## Project Predict: Data Bias and Remediation

**Potential Bias Issue with Project Predict ML Modelling**

*TriCARE Population and Lack of Diversity*

A health monitoring system, such as Project Predict, is arguable more attractive to an older client base where health issues are more common. The five hundred patients are on average above retirement age, although with some younger outliers.

The sample profile presents two obvious problems; the first is that an older Irish population is likely to be less diverse as significant inwards migration to Ireland did not occur until the late 1990s and beyond**[6]**. The second is that patients in their twenties appears to be a small proportion of the TriCARE sample.

An ML model built with this data will skew to an older white Irish demographic (the ratio of males to females is also approximately 2:1). This data imbalance will introduce a bias in the auto ML triage recommendations.

*Understanding/Correcting Triage Recommendations*

The models built by MLHealth must be open to interpretation, so that analysts and stakeholders can understand what features of the Medic sensor data, and the patient medical profile, prompted a specific triage recommendation.

If the MLHealth data scientists elect to use the newer, popular, neural network approaches to building models, then the goal of an explainable (and auditable) recommendation output will be much more complex.

**Proposed Remedies**

*Constant Re-Training of the Triage ML Model*

The Project Predict lifecycle does allow for a process of re-training as new, more ethnically diverse patients are introduced into the system.

However, this may take too long to correct for bias and never really address the fact that younger patients will always be a significant minority in the training dataset.

A policy of artificial data generation is recommended to create ‘non-real’ patients that will temper the ‘real-world’ bias in the MLHealth triage modelling process. However, this must be approached with caution as ML research in other areas has shown that artificial techniques like SMOTE can also introduce distortion into ML datasets**[7]**.

*Adopting Advances in Neural Network ‘Explainability’*

The data scientists in MLHealth should look closely at recent research by Sinanc et al (2021)**[8]** if a neural network model is to be deployed in Project Predict. Such an approach would offer a feature ‘heat map’ that explains the attributes that primarily drove the triage recommendation.

This would allow stakeholders to understand the logic behind an ML decision in the event of a dispute.

## Project Predict: Data Privacy and Security

*Data Security*

There is an informal perception that cloud-based data storage is inherently less secure than ‘on-premise’ data storage. In reality, the major cloud service providers are so conscious of the potential reputational damage of a data breach that their security infrastructures tend to be superior to commercial in-house environments**[9]**.

The key recommendation for Project Predict is that MLHealth and TriCARE must establish that all cloud-based data systems used by DigiHealth and Medic are with one of the major established providers, such as Amazon (AWS) or Microsoft (Azure).

*Authentication Protocols*

Given the involvement of four major software and hardware providers in the delivery of Project Predict, it will be essential that rigid login profiles, and associated privileges, are established as early as possible in the development lifecycle.

No one developer or analyst should have access to sensitive patient information unless absolutely required.

# Data Protection Impact Assessment

## Project Predict: Basis for Lawful Data Processing

Project Predict must adhere to legal requirements, as set out in Article 6 of the 2016 EU GDPR legislation, that each participating patient has given clear consent for their personal data to be collected, processed, and stored by the system**[10]**. The system implemented by all parties (MLHealth/ TriCARE/ DigiHealth) must also inherently observe the legal requirement to process the patient data fairly and transparently, as per the principles in Article 5 of GDPR**[11]**.

A legal audit should take place, both at the start of the project, and at least once during the Year 3 Trial, to ensure that Project Predict remains compliant with the Irish Data Protections Act (2018). This legislation supplements GDPR and Project Predict needs to comply, in particular, with Section 57 of the DPA and ensure that each patient *consents to automated decisions* being made by the MLHealth ML triage process.

## Project Predict: Data Controllers and Processors

**Data Controllers[12]**

*MLHealth*

Company Directors. Joint Controllers.

* Autonomous management responsibility for collection and processing of patient data (sensor and survey).
* Manage direct contracts with patient group, and ultimately profit from Project Predict commercial activity.
* Decide where and how patient data is stored, and for how long.

Data Scientists. Controllers.

May seems to appear as ‘Processors’ given that they build the ML Triage model on behalf of MLHealth management. However;

* The Data Scientist group exercise significant control over patient data in terms of building ML triage models and, more crucially, retraining these models over time to alter triage performance.
* They are interested in the specific outcome of the processing and will re-engineer patient data as required to improve ML model accuracy.

*TriCARE*

Company Directors. Joint Controllers (with MLHealth directors).

* Mirror responsibilities of MLHealth directors and participate in collection of patient data for a common Project Predict purpose.

**Data Processors[12]**

*MLHealth*

Researchers. Processors.

* Follow instructions of MLHealth management to obtain and supplement patient data throughout the Year 3 trial.
* Only MLHealth management/data scientists take operational actions based on this data.

*TriCARE*

Nurses. Processors.

* They do make decisions based on the Medic Sensor data but only within their contracted duties for TriCARE.
* They do not own the data collection process or have any control over which patients to assess.

Triage Team. Processors.

* Take data input from Nurses (or ML model) in terms of a triage recommendations and will decide on appropriate medical interventions. However, this is again done within the confines of an agreed contract with TriCARE, and the Triage Team does not own the data process.

*DigiHealth*

Company Directors. Processors.

* Although this group does have a commercial interest in ensuring their system is managing data flows within Project Predict, this is all within a contract relationship with the Project Predict owners (Controllers).
* DigiHealth management will apply their expertise and judgement in relation to medical data management, but they do not control the lawful basis for collection and processing of that data.
* They do not decide for how long patient data is stored in the system.

Other Employees. Processors.

* All data processing implementation is conducted under contract agreement with a project Controller (TriCARE).

*Medic*

All Employees. Processors.

* All data processing is under instruction by, and contract with, TriCare and DigiHealth.

## Project Predict: Safeguards

There are numerous safeguard recommendations that could be made for Project Predict but two are singled out here in terms of *Purpose* and *Identification*.

*Purpose*

Article 26 of the Irish 2018 Data Protection Act is clear that confidential personal information must not be in appropriately shared by action (or inaction) of a Data Controller**[13]**. Project Predict directors must be certain the processes are in place that no parties share sensitive patient information with external entities for the purposes or Marketing, or other non-authorised activity. Article 27 provides the basis for civil proceeding that any patient within the system would be entitled to peruse in the even of a data ‘leak’. Security protocols must be in place, through user authentication, or database level encryption, to prevent such an eventuality.

*Identification*

Anonymisation is a process that would render it impossible to associate medical data with an identified patient**[14]**. It would have the advantage of allowing Project Predict to sidestep most of the rigours of GDPR and related Irish legislation.

However, the recommendation of this report is that anonymisation is not practical for this system. Year 3 survey data must integrate into the ongoing daily Medic senor data readings for each patient. Therefore, the Project Predict system always retains a means to re-establish patient identify, so no one remains truly anonymous, particularly in the eyes of the data protection laws.

That said, the triage model does not need personal identification details such as name or address. These could be replaced with numerical pseudonyms for most purposes within the project. Mappings of the pseudonyms to real names will be required to process Year 3 survey data but this information should be stored in a more restricted section of the project data store.

## Project Predict: Data Collection and Consent

*Collection of Consent*

It is assumed that any patient supported by the TriCARE organisation at the beginning of the Project Predict lifecycle has already committed to a signed contract agreement. However, given the extended nature of data processing within Project Predict, added terms and conditions must be additional agreed by each individual though a supplementary signed document.

Any patient who does not agree to the terms of Project Predict must not be included in the joint initiatives by MLHealth and TriCARE. Consent is a key tenant of GDPR regulation and Project Predict must respect the will of each individual and their desires for privacy. This data will not be incorporated in the ML modelling process, and these patients must not be approached for the Year 3 Trial survey.

*Withdrawal of Consent*

GSPR legislation is clear on the ‘right to erasure’**[15]** should any patient within the Project Predict system wish to with withdrawn consent to further involvement.

The Joint Controllers within MLHealth and TriCARE must respond as quickly as possible to such a request and provide evidence that the patient data was deleted from the project data stores.

## Project Predict: Ethical and Privacy Risk Matrix

The following ethical and privacy risk matrix has been defined for the lifecycle of Project Predict, based on the risk assessment guide below;

A screenshot of a computer

Description automatically generated with medium confidence

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Ethical/Privacy Risks Identified** | **Severity Risk Score (1-5)** | **Likelihood Risk Score (1-5)** | **Overall Risk Score**  **Sev\*Lik** | **Proposed Solutions(s)** |
| **Does it preserve the autonomy of the human?**  A proportion of patients involved in the Project Predict lifecycle will be elderly and may be uncertain about their data rights, in terms of access and removal. | 4 | 4 | 16 | Project Predict must provide patients with access to an individual (not a website or a chat-bot) who will explain, in straightforward terms, their data rights.  This must cover data access rights throughout the four years of the project and include specific reference to the Year 3 survey. |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Ethical/Privacy Risks Identified** | **Severity Risk Score (1-5)** | **Likelihood Risk Score (1-5)** | **Overall Risk Score**  **Sev\*Lik** | **Proposed Solutions(s)** |
| **Does it preserve or enhance human dignity?**  The MLHealth ML models offer the potential of faster triage response to patient health issues. However, the models are built on a relatively small survey of data, and one that we believe lacks certain elements of diversity. An automated triage decision must not generate a TriCARE response that is potentially inaccurate, and distressing to a patient. | 4 | 3 | 12 | The DPIA flagged concerns with the spread of diversity in the five hundred members of the Project predict sample population.  It is strongly recommended that artificial data ‘profiles’ are added to the modelling data so that triage recommendations are not skewed towards an elderly white male class. |
| **Is processing necessary and proportionate?**  TriCARE is clearly an organisation that has connections with the broader Healthcare industry. Project Predict data could be a valuable commodity for other players in the sector. How can this be addressed? | 5 | 2 | 10 | Given the legal implications of providing PII to external entities for unauthorised use, it is unlikely that MLHealth and/or TriCARE employees would be guilty of such a breach.  However, data management protocols should be a feature of regular staff training throughout the four-year Project Predict lifecycle. |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Ethical/Privacy Risks Identified** | **Severity Risk Score (1-5)** | **Likelihood Risk Score (1-5)** | **Overall Risk Score**  **Sev\*Lik** | **Proposed Solutions(s)** |
| **Does it preserve or enhance human dignity?**  Early intervention to either an emergency condition, or a more gradual deterioration in health, by Medic Sensors, and the data decisions within the TriCARE/MLHealth system, can improve the longer-term well-being of a patient. It can be argued that this is exactly why patients sign up to TriCARE in the first place – they wish to extend quality of life for themselves, if possible. | 3 | 3 | 9 | The ‘risk’ in this instance is that not joining the Project Predict programme will prevent early detection of serious illness. Therefore, once the client has given explicit, written consent to join Project Predict then it is expected that their medical data becomes part of this data ecosystem. |
| **Is processing necessary and proportionate?**  There is a constant recoding of daily patient data, which is then supplemented by the Year 3 Trial survey information. This means that patient data will have to be identifiable at all stages in the Project Predict process. Is it clear who has access to these identifiers and how they are used? | 4 | 2 | 8 | It is expected that DigiHealth will implement a system that linking PII to a specific individual will only be available to a minority of MLHealth and TriCARE staff.  Access will only be available for specific purposes and this access will be tracked in an audit log for later scrutiny should any data breach occur. |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| .**Ethical/Privacy Risks Identified** | **Severity Risk Score (1-5)** | **Likelihood Risk Score (1-5)** | **Overall Risk Score**  **Sev\*Lik** | **Proposed Solutions(s)** |
| **Does it preserve the autonomy of the human?**  Every individual competent adult is entitled to determine how they interpret and react to medical advice. Project Predict should not force triage decisions on a patient or refer this information to an external medical agency (hospital, etc.) without a patient’s consent. | 4 | 2 | 8 | Most patients, by virtue of signing up to the Project Predict programme, are probably willing to accept medically recommended triage decisions.  However, there must be an ‘opt-out’ decision that can be captured after the TriCARE Triage team has assessed the recommendation from the nurse/ML model. Triage information must not leave the Project Predict data perimeter unless explicitly agreed by the patient.  If the patient is unable to provide this triage consent, due to a profoundly serious medical condition, then TriCARE employees should be empowered to act on the patient’s behalf. |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Ethical/Privacy Risks Identified** | **Severity Risk Score (1-5)** | **Likelihood Risk Score (1-5)** | **Overall Risk Score**  **Sev\*Lik** | **Proposed Solutions(s)** |
| **Is processing necessary and proportionate?**  Certain patients may be agreeable to the use of automated Medic sensors but reluctant to take part in the Year 3 trial. Will participation in the survey, and providing that supplementary data, be a pre-requisite to joining Project Predict? | 3 | 2 | 6 | The trial interview sessions are intended to take place either at patient’s home address, or over the phone. Ideally, patients will be briefed well in advance on the purpose of this survey and be willing to get involved.  A patient may be uncomfortable disclosing information to an unknown researcher. An alternate would be to allow the option of completing the survey online, or by post.  The researcher should also ideally be an identified medical professional, who would be qualified to interpret survey responses, if required. |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Ethical/Privacy Risks Identified** | **Severity Risk Score (1-5)** | **Likelihood Risk Score (1-5)** | **Overall Risk Score**  **Sev\*Lik** | **Proposed Solutions(s)** |
| **Does it uphold the common good?**  The USP of the Project Predict services is that it promises a quicker (automated) response when medical sensor data detects a dangerous deterioration in a patient’s health. Project Predict offers a potentially repeatable solution that can market better healthcare for sections of society. | 1 | 5 | 5 | It is a strong belief that initiatives like Project Predict can use data management to delivery societal benefit in early health care issue detection. |
| The Medic sensors are also calibrated to build an ongoing daily history for a patient. This can have the benefit of highlighting emerging conditions and recommending treatment early. This is beneficial to both the patient and to the broader health service by avoiding later (and costly) emergency intervention. | 1 | 5 | 5 | Project Predict operates outside of a hospital or doctor’s surgery setting. This avoids expense for both the individual and the health care service in general. |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Ethical/Privacy Risks Identified** | **Severity Risk Score (1-5)** | **Likelihood Risk Score (1-5)** | **Overall Risk Score**  **Sev\*Lik** | **Proposed Solutions(s)** |
| **Does it preserve the autonomy of the human?**  Patients are consenting to membership of Project Predict, but they are expecting MLHealth and TriCARE to be responsible custodians of their sensitive medical information. | 5 | 1 | 5 | All MLHealth and TriCARE company directors will be required to make a declaration that the understand their data responsibilities under GDPR/Irish Data Protection legislation.  It must also be clear that as Joint Data Controllers they are aware of the legal sanctions that would apply in the event of data breaches.  (This is a core assumption in the establishment of Project Predict, hence the likelihood of an issue is considered ‘remote’). |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Ethical/Privacy Risks Identified** | **Severity Risk Score (1-5)** | **Likelihood Risk Score (1-5)** | **Overall Risk Score**  **Sev\*Lik** | **Proposed Solutions(s)** |
| **Does it uphold the common good?**  If a patient is unwilling to share identifiable personal medical data, they will be excluded from Project Predict. Does this set a precedent for future similar systems? Is PII always an essential ingredient for such systems? | 2 | 2 | 4 | It is likely that clients of future Project Predict systems will understand the benefit of sharing their medical data, given that it offers life enhancing and extending benefits. |

# References

[1] Polkovnikov, I. (2016). Unified Control and Data Flow Diagrams Applied to Software Engineering and other Systems. Retrieved 15 April 2022, from <https://doi.org/10.48550/arXiv.1610.02374>

[2] Saud, D. (2022). What Are The Three Main Goals Of Data Lifecycle Management (DLM)? - Folio3 Dynamics Blog. Retrieved 5 May 2022, from <https://dynamics.folio3.com/blog/what-are-the-three-main-goals-of-data-lifecycle-management-dlm/>

[3] Taylor, M., McNicholas, C., Nicolay, C., Darzi, A., Bell, D., & Reed, J. (2013). Systematic review of the application of the plan–do–study–act method to improve quality in healthcare. *BMJ Quality &Amp; Safety*, *23*(4), 290-298. doi: 10.1136/bmjqs-2013-001862

[4] Krishnan, S., & Shashidhar, N. (2019). eDiscovery Challenges in Healthcare. *International Journal Of Information Security Science*, *8*(2), 30-43. Retrieved from <https://www.researchgate.net/profile/Sundar-Krishnan/publication/335202708_eDiscovery_Challenges_in_Healthcare/links/5d566f4d92851cb74c6facf1/eDiscovery-Challenges-in-Healthcare.pdf>

[5] Orwa, P. (2020). Replication As A Way To Achieve Interoperability in Healthcare. Retrieved 9 May 2022, from <http://erepository.uonbi.ac.ke/bitstream/handle/11295/153117/REPLICATION%20AS%20A%20WAY%20TO%20ACHIEVE%20INTEROPERABILITY%20IN%20HEALTHCA.pdf?sequence=1>

[6] Ireland Immigration Statistics 1960-2022. (2022). Retrieved 9 May 2022, from <https://www.macrotrends.net/countries/IRL/ireland/immigration-statistics>

[7] Priscilla, C., & Prabha, D. (2020). Influence of Optimizing XGBoost to handle Class Imbalance in Credit Card Fraud Detection. *2020 Third International Conference On Smart Systems And Inventive Technology (ICSSIT)*. doi: 10.1109/icssit48917.2020.9214206

[8] Sinanc, D., Demirezen, U., & Sağıroğlu, Ş. (2021). Explainable Credit Card Fraud Detection with Image Conversion. *ADCAIJ: Advances In Distributed Computing And Artificial Intelligence Journal*, *10*(1), 63-76. doi: 10.14201/adcaij20211016376

[9] Clouder, A. (2021). Top 10 Benefits of Cloud Computing. Retrieved 10 May 2022, from <https://www.alibabacloud.com/blog/top-10-benefits-of-cloud-computing_597739>

[10] Art. 6 GDPR - Lawfulness of processing - GDPR.eu. (2022). Retrieved 10 May 2022, from <https://gdpr.eu/article-6-how-to-process-personal-data-legally/>

[11] Art. 5 GDPR - Principles relating to processing of personal data - GDPR.eu. (2022). Retrieved 10 May 2022, from <https://gdpr.eu/article-5-how-to-process-personal-data/?cn-reloaded=1>

[12] Controllers and processors. (2022). Retrieved 10 May 2022, from <https://ico.org.uk/for-organisations/guide-to-data-protection/guide-to-the-general-data-protection-regulation-gdpr/key-definitions/controllers-and-processors/>

[13] Data Protection Act 2018. (2022). Retrieved 10 May 2022, from <https://www.irishstatutebook.ie/eli/2018/act/7/enacted/en/print#sec26>

[14] What is personal data?. (2022). Retrieved 10 May 2022, from <https://ico.org.uk/for-organisations/guide-to-data-protection/guide-to-the-general-data-protection-regulation-gdpr/what-is-personal-data/what-is-personal-data/#pd5>

[15] The right to erasure (Articles 17 & 19 of the GDPR) | Data Protection Commission. (2022). Retrieved 10 May 2022, from <https://www.dataprotection.ie/en/individuals/know-your-rights/right-erasure-articles-17-19-gdpr#:~:text=This%20is%20also%20known%20as,it%20was%20collected%20or%20processed>.