|  |  |
| --- | --- |
|  | |
| 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  15/4/2022 |  |
|  |  |

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 (Wk10) 5

2.3 Project Predict: Data Bias and Remediation (Wk9) 7

2.4 Project Predict: Data Privacy and Security (Wk11) 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 10

3.3 Project Predict: Safeguards 10

3.4 Project Predict: Data Collection and Consent 10

3.5 Project Predict: Ethical and Privacy Risk Matrix 11

4 Conclusions 12

5 References 13

# Introduction

## Purpose of Report

The..

## The Data Management Plan

The..

## The Data Protection Impact Assessment

The..

# 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 act as a proxy for assessments carried out by nurses.

## Project Predict: Data Quality Issues (Wk10)

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 is being treated with the upmost respect. (Taylor et al., 2013) provide excellent guidelines in how to apply this in a healthcare setting, such as Project Predict**[3]**.

*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, outside of possible medical ranges. The potential challenge with such Medical Internet of Things (IoT) is one of the challenges discussed in tr 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 different 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 (Wk9)

**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 500 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 20s appears to be a very small proportion of the TriCARE sample.

An ML model built with this data will skew somewhat 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 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 to 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 offers a feature ‘heat map’ that explains the attributes that primarily drove the triage recommendation.

This would allow stakeholder to understand the local behind an ML decision and offer possible means of redress in the event of a dispute.

## Project Predict: Data Privacy and Security (Wk11)

*Data Security*

There is a somewhat 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 infrastructure tends to be superior to many 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

Ipso Lorem…

Ipso Lorem…

## Project Predict: Data Controllers and Processors

Ipso Lorem…

Ipso Lorem…

## Project Predict: Safeguards

Ipso Lorem…

Ipso Lorem…

## Project Predict: Data Collection and Consent

Ipso Lorem…

Ipso Lorem…

## Project Predict: Ethical and Privacy Risk Matrix

Ipso Lorem…

Ipso Lorem…

<Diagram>

# Conclusions

One could ipso lorem…

.

# 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