Project Acronym:	Nutrishield
Grant Agreement number:	818110 (H2020-SFS-2018-IA)
Project Full Title:	Fact-based personalised nutrition for the young

DELIVERABLE

D2.5 – Data fusion & local application requirements

Dissemination level	PU - Public
Type of Document	Report
Contractual date of delivery	30/06/2019
Deliverable Leader	Cranfield University
Status & version	Final 1.0 – 05/07/2019
WP responsible	WP2 (WP Leader RU)
Keywords:	Machine Learning – Data Science – Personalised Medicine Algorithm – Mathematical Modelling – Classification - Regression

Deliverable Leader:	Cranfield University
Contributors:	Fady Mohareb (CU) – Corentin Molitor (CU) – Simona Cristescu (RU) – Julia Kuligowski (UV) - Intrasoft
Reviewers:	M. Vasileiadis (ALPES)
Approved by:	M. Vasileiadis (ALPES)

Document History			
Version	**Date**	**Contributor(s)**	**Description**
v0.1	26/06/2019	Fady Mohareb	Draft
v0.2	27/06/2019	Corentin Molitor	Draft
v.03	27/06/2019	Fady Mohareb	Draft
v.041	04/07/2019	Miltos Vasileiadis	Draft
V1	04/07/2019	Fady Mohareb	Final

This document is part of a project that has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 818110. It is the property of the NUTRISHIELD consortium and shall not be distributed or reproduced without the formal approval of the NUTRISHIELD Management Committee. The content of this report reflects only the authors' view. EC is not responsible for any use that may be made of the information it contains.

Executive Summary

This deliverable is related to the WP2 task entitled "*T.3 Requirements regarding fusing of biomarker data & personalized nutrition algorithm*" (Task Leader – Cranfield University – CU). This task aims to create a list of requirements, prior to the project clinical studies, for the personalized nutrition machine learning algorithm to act as a decision support system to the Nutrishield Platform. The personalised nutrition algorithm will be based on a series of classification and regression models, which take into account multiple dimensions of heterogeneous clinical, nutritional, genotypic, and metabolic parameters, measured/recorded throughout the course of this project. Due to the heterogeneous nature of the input variables for the machine learning models, a novel data fusion approach will be developed in order to integrate these system-level, multi-dimensional knowledge level into a single layer for the prediction algorithm. The data fusion methodology is beyond the scope of this deliverable and will be discussed in detail in D7.5. Therefore, this report will only focus on the actual data input requirements.

Generally speaking, the main model training datasets will include:

- human milk analysis,
- microbiome analysis,
- blood and urine analysis, and
- genotypic data (SNPs and InDels profiling) of the clinical subjects.

Additional metadata that would potentially improve the prediction accuracy of the models include the current dietary profile for the patient, as well as additional epidemiological parameters (age, sex, ethnic origin, etc.). Additional output from the novel sensors developed as part of this project (WP4) will also be assessed via a separate machine learning prediction layers and comparatively validated against other established methods in order to evaluate their suitability as part of the Nutrishield nutrition platform.

Table of Contents

1. Requirement regarding the biomarkers data fusion algorithm 5

1.1 Data format 5
1.1.1 Patient genotyping 5
1.1.2 Metabolic profiles 5
1.1.3 Data sharing and storage 6

1.2 Personalised nutrition algorithm requirement 6
1.2.1 Background 6
1.2.2 The algorithm handling of the different datasets. 6
1.2.3 General guidelines 7
1.2.4 Nutritional rules 7
1.2.5 Model development 8
1.2.6 Model training and optimisation 8

Definitions, Acronyms and Abbreviations

Acronym	Title
LDA	Linear Discriminant Analysis
PLS-DA	Partial Least Square Discriminant Analysis
SVM	Support Vector Machine
RF	Random Forest
LR	Linear Regression
PLSR	Partial Least Square Regression
PCR	Principal Component Regression
kNN	K-Nearest Neighbours
VCF	Variant Call Format

1 Requirement regarding the biomarkers data fusion algorithm

1.1 Data format

One of the main requirements for an effective data sharing strategy between different partners involved in collection and analysis of clinical data is to ensure that cross-compatibility is established between the machine-learning model and the raw data input from various analytical platforms and sensors deployed during the study. Due to the fact that a large number of devices will be used, including novel sensors, a suitable data exchange protocol may not always be already in place. Furthermore, developing a novel data exchange protocol for every entry will be time-consuming and beyond the scope of this project. Following discussion between all partners involved, both during the 6M meeting and during teleconferences, It was agreed that all measurement/biomarker data will be shared, whenever possible, in its original raw and unprocessed form, in CSV format. This will also be the case for multispectral and hyperspectral data. For this purpose, spectral data will be converted internally to intensities measurements as table columns before being sent to the algorithm for prediction with the sample names (i.e. patient anonymized ID) as the first column. However, for sequencing data related to genotyping and microbiome analysis, high throughput data will be used in its original format (fastq, fasta, etc.) since it is still a textual form that can be parsed with minimal intervention from the modelling side as described below.

1.1.1 Patient genotyping

We will perform genotyping by sequencing followed by variant calling for our patient subject using Illumina HiSeq paired-end reads. Raw sequence reads will be generated in FASTQ format, while assembled genome sequence will be developed in FASTA format. Genome functional annotation will be generated in TSV and GFF3 formats. Variant calling files will be generated in VCF format (Expected to be ~1Gb file per patient, our clinical study will include ~120 individuals in total). All patient details and metadata will be anonymised before being sent to Cranfield as per the partner clinical regulation.

The paired-end reads obtained with the HiSeq platform will be aligned against the human reference genome GRch38 (Schneider et al., 2016) and variant calling performed using GATK best practices (DePristo et al., 2011). Variants will also be annotated in order to find the most relevant variants for each clinical study.

1.1.2 Metabolic profiles

Additional clinical data and metabolic profile will be mostly in a tabular format (Anonymised patient IDs as rows and biomarkers are columns) in MS Excel or CSV format (relatively small files of <500 kb each).

1.1.3 Data sharing and storage

All Nutrishield experimental results (genotypes and clinical/metabolic profiles/ epidemiological data) will be hosted on a RAID expansion server that provides 120 TB of raw storage capacity. The expansion server will be added to our existing HPC storage facility hosted at Cranfield University. The storage facility is equipped with enterprise-grade storage hard drives configured on a RAID10 system that provides data protection during multiple drives failure. Data stored on the RAID expansion server will be supported for a period of 10 years beyond the course of this project.

1.2 Personalised nutrition algorithm requirement

1.2.1 Background

Requirements as to how the personalised nutrition algorithm will take into account the biomarker data to give dietary advice. The final algorithm will deliver personalised nutrition advice based on biomarkers and their time-dependent evolution. Generating the advice will require the integration of mechanisms identified from the analysis of the data collected during the study as well as established nutritional rules. The set of nutritional rules integrated in the final algorithm will need to be identified and given a ranking based on priority. A process that will be performed in collaboration with the rest of the Nutrishield consortium. Next, requirements of the data-driven portion of the algorithm needs to be determined. In algorithm design there is a general trade-off between completeness and complexity (Handelman et al., 2019).

Algorithms that attempt to be too complete will become too complex, and those that remain too simple will lack completeness. The level of desired complexity needs to be defined. There is a large variety of algorithm for the analysis of Big Data and its use in making predictions. A list of elements that could be integrated in the final solution needs to be set up, using both existing and custom design elements

1.2.2 The algorithm handling of the different datasets.

The objective is to develop a machine-learning algorithm that will detect any discrepancies in the metabolites levels of a patient against healthy levels. It will also provide personalised nutritional advice (based on the EU standards) to attempt to correct these discrepancies.

The first step will be to identify the important metabolites / biomarkers resulting from an unhealthy diet. This will be done by using historical data from the Nutrishield partner. Algorithms will be trained on these datasets and the most important variables for the classifications will be determined. The table below describe how each dataset measured during the clinical studies will be used by the algorithm.

Data type	Use
Genotyping	● Compare the list of variants obtained to a curated list of variants associated with Diabetes & Obesity. This will create a risk score for the patient for each disease. ● Depending on the cohort sizes, there is a possibility of discovering new variants linked to these diseases.
Microbiome	● Compare the diversity of the microbiome (number of different species) and the species lists to link it with response to different food. A diverse microbiome has been associated to better health
Metabolites / vitamins / biomarkers	● Assessment of the metabolite levels against healthy levels.
Questionnaires	● Classification from the dietary habits can be performed to try to see differences in metabolites / microbiomes between different “dietary groups” (if exist). ● The Mediet score can also be calculated and advice given to bring this score to an optimal level. ● Sociodemographic information will be used to assess if it has an impact on the diet / health.
Breath	● Given as m/z spectra. ● Volatile Organic Compounds will be measured and assessed against typical concentrations. ● Historical data has already been processed.

1.2.3 General guidelines

We should avoid subjective values, as “low”, “Medium”, “high” as these can mean different things depending on the person entering the values. Objective measurements will result in a better algorithm. Project partners need to ensure agreement upon on a standard for healthy levels of vitamins / nutrients intake etc.

1.2.4 Nutritional rules

The rules will be based, among other sources, on the “EU Register on nutrition and health claims” (http://ec.europa.eu/food/safety/labelling_nutrition/claims/register/public/?event=search and https://ec.europa.eu/food/safety/labelling_nutrition/claims/nutrition_claims_en). Only claims with a validated status ("authorised") will be used, since they are supported by scientific evidence.

1.2.5 Model development

The models will be developed based on the clinical data collected as part of Phase I – the observational phase of clinical trials (See Deliverable 2.7 for more information). Briefly, children and lactating mothers will be recruited and their diet and biomarkers will be assessed over the first four month of the trial. Using this data, a variety of machine learning methods will be developed as part of the personalized nutrition algorithm. This includes the following classification and regression algorithms Linear Discriminant Analysis (LDA), Partial Least Square Discriminant Analysis (PLSDA), Support Vector Machine (SVM), Random Forest (RF), Linear Regression (LR), Partial Least Square Regression (PLSR), Principal Component Regression (PCR) and K-Nearest Neighbours (KNN).

LDA and PLSDA are classification methods used for multivariate dataset containing correlated variable. The first step of the algorithm is to lessen the number of variables to solve problem by minimizing data variation in the same class and increasing the separation between classes. RF grows many decision trees. Each tree is constructed using a different subset from the original one. Trees predictions are aggregated and the result remains from the majority. Thus, RF avoid overfitting data. SVM is a non-linear classifier which working principle is to draw a hyperplane to separate data. The latter is determined by maximizing the distance between the closest points. It is effective for classification and regression even when number of dimensions is greater of number of sample. KNN is a non-parametric method gathering samples in group of K closest individual. Groups are defined calculating the Euclidean distance between all the points. For classification, the output is the class of the majority of samples. For regression, the output is the average bacterial count.
LR fines the best fitting straight line through the data points. The line represents the predicted values. Best fitting line is the one minimizing the sum of squared errors prediction
PCR and PLSR are two regression algorithms designed to work with data suffering from multicollinearity. Before performing linear regression, dataset are reduced using principal component analysis (PCA) for the first method and PLS algorithm for the second one.

1.2.6 Model training and optimisation

The model will be trained using the data collected during the observational period, and optimised against the clinicians' assessment of their profile based on the collected measurement. The datasets will be split in two groups, the training dataset (~70-80% of the samples), which will be used to train the algorithm and the test dataset (~20-30% of the samples), which will be used to validate the algorithm. Indeed, the performance of each one of the models will calculated by comparing the predicted values of the model with the unseen actual values of the testing dataset. This will result in a series of classification models for factorial (classes) separation (e.g. normal vs. abnormal – diseased vs. control etc.). The model accuracy will be simply calculated as a function of the total correct classed test samples per the total testing dataset (Equation 1). Additionally, a series of regression models to predict a value based on multivariate data input (e.g. predict a metabolite level based on breath profile analysis. For this purpose, the model accuracy will be calculated based on s the root-mean-square error (RMSE). The models with the higher accuracy or the lowest RMSE were selected as the best ones. The accuracy is an indicator of the number of predicted values that match their correspondent observed values so, the closer that is to 1, the better the model is (Equation 1).

$Accuracy = \frac{samples\ correctly\ predicted}{total\ number\ of\ samples}\ x\ 100$	**2**

On the other hand, RMSE quantifies the difference between predicted and observed values, if the difference is small, then RMSE is a positive number close to 0 (Equation 2).

$RMSE = \sqrt{\frac{\sum (predicted - observed)^2}{n}}$	**3**

The normalised RMSE (NRMSE) was also calculated. RMSE only can take as maximum value the difference between the highest and the lowest values of the dependent variable (DV), which is the actual TVC. Therefore, dividing RMSE between the difference of the maximum DV and the minimum DV will give a percentage of error that makes easier the comparison among different models (Equation 3).

$NRMSE = \frac{RMSE}{(DV) - (DV)}$	**4**

2 References

DePristo, M. A., Banks, E., Poplin, R. E., Garimella, K. V., Maguire, J. R., Hartl, C., … Daly, M. J. (2011). A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nature Genetics*, *43*(5), 491. https://doi.org/10.1038/NG.806

Handelman, G. S., Kok, H. K., Chandra, R. V., Razavi, A. H., Huang, S., Brooks, M., … Asadi, H. (2019). Peering Into the Black Box of Artificial Intelligence: Evaluation Metrics of Machine Learning Methods. *American Journal of Roentgenology*, *212*(1), 38–43. https://doi.org/10.2214/AJR.18.20224

Schneider, V. A., Graves-Lindsay, T., Howe, K., Bouk, N., Chen, H.-C., Kitts, P. A., … Church, D. M. (2016). Evaluation of GRCh38 and de novo haploid genome assemblies demonstrates the enduring quality of the reference assembly. *BioRxiv*, 072116. https://doi.org/10.1101/072116