Project title	Artificial intelligence and the personalized prevention and management of chronic conditions		
Project acronym	WARIFA		
Project number	101017385		
Call	Digital transformation in Health and Care	**Call ID**	H2020-SC1-DTH-2020-1
Topic	Personalised early risk prediction, prevention and intervention based on Artificial Intelligence and Big Data technologies	**Topic ID**	SC1-DTH-02-2020
Funding scheme	Research and Innovation Action		
Project start date	01/01/2021	**Duration**	48 months

D1.14 – DATA MANAGEMENT PLAN

Due date	30.06.2021	**Delivery date**	30.06.2021
Work package	WP1 Project Management		
Responsible Author(s)	Thomas Schopf and Conceição Granja Bartnæs		
Contributor(s)	All partners		
Version	1.0		

DISSEMINATION LEVEL

Please select only one option according to the GA			
☒	PU: Public	☐	PP: Restricted to other program participants
☐	RE: Restricted to a group specified by the consortium	☐	CO: Confidential, only for members of the consortium

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 101017385

VERSION AND AMENDMENTS HISTORY

Version	Date (MM/DD/YYYY)	Created/Amended by	Changes
0.0	22.04.2021	Merethe Drivdal	Creation of document
0.1	05.05.2021	Merethe Drivdal	Update on all chapters
0.2	08.05.2021	Conceição Bartnæs	Update
0.3	27.05.2021	Eva Henriksen	DPO Input
0.4	27.05.2021	Thomas Schopf	Update
0.5	26.06.2021	Conceição Bartnæs	Update
0.6	26.06.2021	Thomas Schopf	Review

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 101017385

TABLE OF CONTENTS

1 INTRODUCTION 5
2 DATA SUMMARY 5
3 DATA QUALITY MANAGEMENT 6
3.1 DATA QUALITY MANAGEMENT – DATA ACQUISITION PHASE 6
3.1.1 Data types in WARIFA System 6
3.1.2 Data acquisition approaches in WARIFA System 7
3.1.3 Data Quality Management Methodology/Approaches 7
3.2 DEFINITIONS 9
3.2.1 Data Quality Dimensions 9
3.2.2 Generic Data Requirement 10
3.2.3 Data Quality Problem Types 11
4 WARIFA DATA MANAGEMENT PLAN 15

LIST OF FIGURES

Figure 1 Different phases of data processing throughout the system [reprinted from (Fürber, 2016)] 6

LIST OF TABLES

Table 1 List of the datasets collected by WARIFA WPs and description of the purpose for which they will be used. 5
Table 2 Datatypes in WARIFA and their acquisition approach, data quality dimensions considered (refer to Table 2 for further details), generic data requirements (refer to Table 3 for further details), and their associated problem (refer to Table 4 for further details). 7
Table 3 Data quality dimensions, their categories, and definitions (Goodhue, 1995, DeLone and McLean, 1992, Wang and Strong, 1996, Wand and Wang, 1996, Fürber, 2016, Cai and Zhu, 2015). 9
Table 4 General data requirements and their associated problems (Fürber, 2016). 10
Table 5 Different types of data quality problems (Fürber, 2016). 11

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 101017385

List of Abbreviations

Abbreviation	Meaning
AI	Artificial Intelligence
CCs	Chronic Conditions
DMP	Data Management Plan
DoA	Description of Action
GDPR	General Data Protection Regulation
TIQM	Total Information Quality Management
TDQM	Total Data Quality Management
TRL	Technology Readiness Levels
TSD	Services for sensitive data (Tjenester for Sensitive Data)
WP	Work Package

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 101017385

1 INTRODUCTION

The WARIFA project aims to facilitate personalised early risk prediction, prevention and intervention based on Artificial Intelligence (AI) and Big Data technologies. We want to explore how AI-based mobile applications may be used as a tool for individual lifestyle changes. This includes the use of mobile applications to analyse and estimate individual risk, correlate it with the community risk profile, provide evidence-based and personalized advice together with prompts for preventive lifestyle changes. The aim is to empower citizens to self-monitor the implementation of risk-reducing lifestyle changes. WARIFA will develop an AI-based system with the aim to help prevent chronic conditions (CCs) for all citizens.

By combining ubiquitous data from the user`s environment with user-generated data, AI algorithms can process the most relevant data in the appropriate context and then provide the tools for personalized advice resulting in more specific preventive interventions. AI and big data technologies have the potential to address these challenges by analysing risk levels and providing citizens with tailor-made advice according to the individual risk level.

To achieve this objective it is necessary to combine ubiquitous data and personal user-generated data, and to combine interdisciplinary efforts from clinical, technical, and sociology background, in order for the WARIFA prototype to reach TRL 6-7 by the end of the project period. The development of the WARIFA the system will be iterative with respect to design/development /testing/feedback-adjustments. Human subjects will be involved in several steps of the project.

The Data Management Plan provides an overview of the datasets collected and generated by the project and to define WARIFA data management policy that is used with regard to these datasets.. This deliverable describes the general policy and approach to data management in WARIFA that handles data management related issues on the administrative and technical level..

The Data Management Plan shall be updated, as necessary, during the project, and is kept available for all WARIFA project members on the chosen platform for project interaction, Microsoft Teams.

2 DATA SUMMARY

The purpose of the data collected within WARIFA is to perform the project activities as described in the Description of Action (DoA), as well as to validate the AI algorithms developed by the project.

Table 1 List of the datasets collected by WARIFA WPs and description of the purpose for which they will be used.

Dataset Name	Purpose
WP2 Mapping	- Identify subpopulations of vulnerable groups and assess their needs, especially in a geographical and cultural context. - Reach a consensus on how AI-based mHealth technology, especially smartphones, may improve the integration of health care services both at the individual and community level with a special emphasis on preventive measures
WP3Sensor data WP3 Ubiquitous data WP3 Registry data	- Identify the most relevant features to enhance existing risk prediction tools that can be found in the clinical setting of CCs

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 101017385

	- Propose new techniques for descriptive and advanced analysis to obtain more effective and interpretable solutions - Identify behavioural patterns and facilitate the health decision support process - Build a probabilistic model that describes quantitatively the causal relationship between indiviual risk factors
WP8 Surveys and interviews	- Measure stakeholder characteristics, e.g., their interest, attitude, influence and knowledge relevant for the project - Identification of policy recommendations

All data collected during WARIFA activities will handled following the procedures described in D1.11 for sensitive data.

3 DATA QUALITY MANAGEMENT

In the literature, there exist several definitions of data quality based on the perspectives and dimensions under considerations (Fürber, 2016, Lee et al., 2002, Wand and Wang, 1996, Goodhue, 1995, Wang and Strong, 1996, Cai and Zhu, 2015, Strong et al., 1997) and, for example, in general, can be defined as "Data are of high quality if they are fit for their intended uses in operations, decision making, and planning. Data are fit for use if they are free of defects and possess desired features." (Fürber, 2016, Redman, 2001). Data quality problem could arise at different data phases of the information system, which can be categorized as data acquisition, data usage, and data retirement (Fürber, 2016), as shown in Figure 1. In this regard, the data quality management methodologies that will be exploited during WARIFA's data acquisition phase are discussed below.

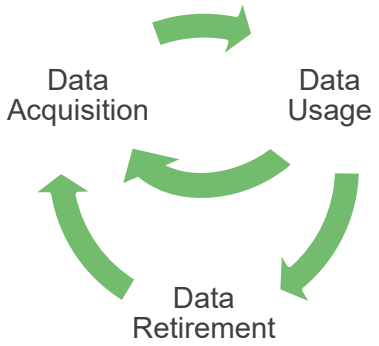

Figure 1 Different phases of data processing throughout the system [reprinted from (Fürber, 2016)]

3.1 DATA QUALITY MANAGEMENT – DATA ACQUISITION PHASE

3.1.1 Data types in WARIFA System

The WARIFA system requires a quality dataset to function properly, and a poor-quality dataset could result in unexpected output and degraded performance. In particular, this emanates from the machine learning model requirements of quality data for satisfactory performance. In this regard, the expected datasets are a list of health parameters based on clinicians' consensus for risk predictions.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 101017385

The data types can be categorized roughly into Demographic/General information, Self-collected data, Clinical and lab data, Medical history, current medication, and Community registry data.

3.1.2 Data acquisition approaches in WARIFA System

In practical settings, the data acquisition will be carried out in two ways; through the WARIFA app (generating new individual data) and retrieving from existing data sources, and storing it into the destination server. The WARIFA app can generate the individual data manually, e.g., forms and questionaries, and automatically, e.g. sensors. Data retrieval from the existing data registries is planned to be carried out using manual and automated data migration and extraction tools and algorithms.

3.1.3 Data Quality Management Methodology/Approaches

There are various types of data quality problems that can arise during the data acquisition phase. From the literature, there are several types of quality data management methodologies or approaches suggested to mitigate and preserve the data quality (Fürber, 2016, Redman, 2001, Wang and Strong, 1996, Lee et al., 2002, Cai and Zhu, 2015). Those methodologies mainly treat structured data, and a few of them treat semi- and non-structured data (Francisco et al., 2017). Choosing the correct methodology to be used is a significant challenge faced by different organizations to systematically address the data quality problems and issues. Two of the well-known data quality management methodologies are “Total information quality management (TIQM)” and “Total data quality management (TDQM)” (Fürber, 2016, Francisco et al., 2017). Due to its practical significance and practical experience-driven nature, TIQM and mainly “reengineer and cleanse data” subtask will be used as overall methodologies (Francisco et al., 2017, Fürber, 2016). Mainly, the following specific task will be carried out (Fürber, 2016):

- • Data quality monitoring provides a mechanism to continuously check the quality of the data and identify data instances with data quality problems,
- • Data quality assessment provides quality data assurances of the various data sources used and mainly the existing public registry data,
- • Data cleansing functionalities provides a mechanism to eliminate data quality problems through formatting and cleaning the datasets to the desired state, and
- • Data constraints provide a mechanism to safeguard data quality during data collection and mainly through data quality rules that can be automatically applied to avoid the generation of data quality problems.

You will find the summary of the datatypes assessed and used in the WARIFA project below, and the suggested associated data acquisition approaches, together with the different data quality dimensions, data requirements, and possible data quality problems, see Table 1.

Table 2 Datatypes in WARIFA and their acquisition approach, data quality dimensions considered (refer to Table 2 for further details), generic data requirements (refer to Table 3 for further details), and their associated problem (refer to Table 4 for further details).

Data types	Data acquisition approach	Data Quality Dimensions	Generic Data Requirement	Data Quality Problem Types

This project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No 101017385

Demographic/ General information	WARIFA app, questionaries, and forms	• Intrinsic - Believability - Accuracy - Objectivity - Reputation • Contextual - Timeliness - Completeness - Relevancy - Appropriate amount of data • Representational - Interpretability - Ease of understanding - Representational consistency - Concise representation • Accessibility - Accessibility - Access security	Property completeness requirements	Missing values, conditionally missing values
	Automatic retrieval, i.e. weather information			
Self-collected data	WARIFA app, automatic and manual retrieval from sensors or storage		Syntactic requirements	Syntax violations, misspelling / mistyping errors, Embedded values, imprecise values
Clinical data	WARIFA app, i.e. questionaries and forms		Legal value requirements	Syntax violations, misspelling / mistyping errors, embedded values, imprecise values, false values, meaningless values, misfielded values
Medical history	WARIFA app, i.e. questionaries and forms		Legal value range requirements	Out of range values, meaningless values, false values
			Illegal value requirements	False values, meaningless values, misspelling / mistyping errors
Current medication	WARIFA app, i.e. questionaries and forms		Functional dependency requirements	False values, referential integrity violations, incorrect references, contradictory relationships
Community registry data	Automatic or manual retrieval,		Unique value requirements	Unique value violations
			Duplicate instance identification requirements	Inconsistent duplicates, approximate duplicates
			Update requirements	Outdated values
			Expiration requirements	Outdated values

More details about the different dimensions, requirements, and problems related to the data types considered in WARIFA are defined and presented below and in Table 2-4.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 101017385

3.2 DEFINITIONS

The following definitions provide the basis for the type of data quality dimensions being considered during the data acquisition phase, the generic data requirements expected of the data during the collection, and also the various problems that could arise in the process and degrade the quality of data being collected.

3.2.1 Data Quality Dimensions

Assessment of data quality relies upon the quality dimensions considered for measurement. In this regard, generally, there are several data quality dimensions (Zmud, 1978, Jarke and Vassiliou, 1997, DeLone and McLean, 1992, Goodhue, 1995, Ballou and Pazer, 1985, Wand and Wang, 1996, Cai and Zhu, 2015), and our data quality assessment methodology relies on the following dimensions of the data as shown in Table 2 below; Intrinsic, Contextual, Representational, and Accessibility dimensions. More detailed analysis and description of those dimensions can be found in (Fürber, 2016, Wang and Strong, 1996, Zmud, 1978, Jarke and Vassiliou, 1997).

Table 3 Data quality dimensions, their categories, and definitions (Goodhue, 1995, DeLone and McLean, 1992, Wang and Strong, 1996, Wand and Wang, 1996, Fürber, 2016, Cai and Zhu, 2015).

Category	Dimension	Definition
Intrinsic	Believability	"The extent to which data are accepted or regarded as true, real and credible."
	Accuracy	"The extent to which data are correct, reliable, and certified free of error."
	Objectivity	"The extent to which data are unbiased (unprejudiced) and impartial."
	Reputation	"The extent to which data are trusted or highly regarded in terms of their source or content."
Contextual	Timeliness	"The extent to which the age of the data is appropriate for the task at hand."
	Completeness	"The extent to which data are of sufficient depth, breadth, and scope for the task at hand."
	Relevancy	"The extent to which data are applicable and helpful for the task at hand."
	Appropriate amount of data	"The extent to which the quantity and volume of available data are appropriate."
Representational	Interpretability	"The extent to which data are inappropriate language and units and the data definitions are clear."
	Ease of understanding	"The extent to which data are clear without ambiguity and easily comprehended."
	Representational consistency	"The extent to which data are always presented in the same format and are compatible with previous data."
	Concise representation	"The extent to which data are compactly represented without being overwhelming (i.e., brief in presentation, yet complete and to the point)."
Accessibility	Accessibility	"The extent to which data are available or easily and quickly retrievable."
	Access security	"The extent to which access to data can be restricted and hence kept secure."

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 101017385

3.2.2 Generic Data Requirement

Data quality requirements are mainly dependent on the type of information system being designed and developed. In the process, every single data type could have a specific requirement needed to be met for the task at hand, and however, in general, there are general requirements expected from any dataset to be of high quality, as shown in Table 3 below, and more detailed description of those requirements can be found in (Fürber, 2016, Cai and Zhu, 2015). The table below depicts the general data requirements for quality data and the associated problem that could degrade the quality of data being collected.

Table 4 General data requirements and their associated problems (Fürber, 2016).

Data Requirement	Data Quality Problem Type	Example
Property completeness requirements	Missing values, conditionally missing values	"Attributes latitude and longitude must have values in table Location to be able to navigate to each location."
Syntactic requirements	Syntax violations, misspelling / mistyping errors, Embedded values, imprecise values	"The attribute country-name must only contain letters and no numbers."
Legal value requirements	Syntax violations, misspelling / mistyping errors, embedded values, imprecise values, false values, meaningless values, misfielded values	"The attribute gender must only contain the values "male", "female", "m", or "f"."
Legal value range Requirements	Out of range values, meaningless values, false values	"The attribute blood glucose levels must only contain non-negative values."
Illegal value requirements	False values, meaningless values, misspelling / mistyping errors	"The attribute gender may never contain the value "mail"."
Functional dependency requirements	False values, referential integrity violations, incorrect references, contradictory relationships	"The attribute city is always dependent on the value for the attribute country since certain city names only exist in certain countries."
Unique value requirements	Unique value violations	"Each value for the attribute ISBN in instances of table Book may not occur more than once."
Duplicate instance identification requirements	Inconsistent duplicates, approximate duplicates	"Instances with the same value for the attribute ISBN and instances with texts that have a similarity greater than 90 % can be considered duplicates."
Update requirements	Outdated values	"Instances of the table Weather are outdated if their

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 101017385

		last modification is more than two years ago.”
Expiration requirements	Outdated values	“Instances of the table CurrentWeather are outdated if their value for the attribute valid until is prior to the current date and time.”

3.2.3 Data Quality Problem Types

There are various kinds of data quality problem associated with the data acquisition phase (Fürber, 2016, Cai and Zhu, 2015, Strong et al., 1997, Pipino et al., 2002, Maydanxhik, 2007), and these mainly can be categorized into Quality Problems of Attribute Values (Single source), Multi-Attribute Quality Problems (Single source, and Integration specific (multiple sources)), Quality Problems of Data Models, and Common Linguistic Problems. A detailed description of these problems is given in Table 4 below, and more details can be found in (Fürber, 2016).

Table 5 Different types of data quality problems (Fürber, 2016).

Problem type	Issues
Quality Problems of Attribute Values Single source	**Invalid characters:** A character that is not expected within a data (Fürber, 2016). For example, consider a questionnaire asking a *piece of demographic information* and particularly a gender value, i.e. male/female, and if a user provides a numeric value within a gender data field, it is called an *invalid character problem*.
	Character alignment violation: This category is defined according to predefined syntax rules and happens mainly because of placing a substring or character in a wrong position (Fürber, 2016). For example, consider a field asking a user to provide *a piece of demographic information* particularly his/her age, and the syntax rule on the questionnaire form dictates to be filled as “*MM/DD/YYYY*“, where M represents the index position for numerical month values, D for numerical day values, and Y for numerical year values. However, if the user provides a value as “30.06.01”, this is invalid according to the syntax rules and called a character alignment violation. Apart from this, misspelling and transpositions are also included in this category.
	Missing values: This is an empty or NULL value supplied by the user for a field that requires a value (Fürber, 2016). For example, for a field asking the hair colour of a user, if the user supplied either blank whitespace or a default value, then it is considered a missing value problem.
	False values: A value that follows the correct syntax rules and holds possible values, however, it doesn’t represent the right state of the underlying entity (Fürber, 2016). For example, the attribute “age” and “skin colour” of a user “Mark peter” has a value “34” and “white”, but Mark’s real age and skin colour are 40 and black.
	Meaningless values: This is a value that doesn’t have meaning and lacks proper interpretability due to the lack of a corresponding real-world entity (Fürber, 2016). For example, the attribute value “name” holds a value equal to “ABC XYZ”.
	Outdated values: A value is outdated if the values of an attribute are obsolete (Fürber, 2016). For example, consider a weather information

This project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No 101017385

	source updated a long time ago and could not be used for current weather information. Accessing any information from such types of sources can affect the decision process that relies on such types of data sources.
	Embedded values: This type of substring value provides additional information and any such value that doesn't fit the intention of the attribute are known as invalid substrings (Fürber, 2016). This type of problem mainly arises in connection to the prefix and postfix added to a data type and for example, the attribute name could hold the title of the user, i.e. "Dr. Mark Johnson". Moreover, the attribute blood glucose level could hold a unit and a value, "6 mmol/l or 108 mg/dl".
	Out-of-range values: A value is declared to be out of range if the value resides outside of the legal range or predefined interval (Fürber, 2016). For example, the attributes "blood glucose levels", "insulin intake" and "carbohydrate consumption" must not contain negative values. Moreover, the attribute "insulin intake" and "carbohydrate consumption" can take reasonable values and, however, values such as "80 units of insulin" and "1000 grams of carbohydrate" depicts an out of range values.
	Imprecise values: A value is declared imprecise if the value of the attribute is ambiguous and poses difficulty to map precisely the corresponding real-world state. This kind of problem mainly occurs within textual attributes and could also arise from homonyms, an attribute containing a value with more than one specific meaning. For example, this can occur in abbreviations and cryptic values (Fürber, 2016).
	Unique value violation: A unique value violation is declared for attributes that mustn't contain more than one value. This kind of attribute is mainly related to values that are meant to serve as identifiers for entities for cross-reference (Fürber, 2016). For example, the attributes "hair colour" and "skin colour" of a user should hold unique values for each tuple.
	Cardinality constraint violation: A cardinality constraint violation is declared if the user-supplied input exceeded the allowed number of values per a single entity (Fürber, 2016). For example, a user can't supply more than one single input for the attribute "date_of_birth".
Multi-Attribute Quality Problems ***Single source and Integration specific (multiple sources)***	**Functional dependency violation (Single source)**: A functional dependency is defined as any inter and intra dependency that exists between two or more attribute values within a tuple, a subset of tuples, and even data sources (Fürber, 2016). For example, consider the geographical location estimation with the attributes "ZipCode" and "Country" and if the user-supplied a value "9018" and "Norway", then the city must be "Tromsø".
	Referential integrity violation (Single source): "If an attribute of one entity comprises values that refer to tuples of another entity, we can call the values of the first attribute "foreign keys". In case of a referential integrity violation, a foreign key value does not have a matching value in the referenced entity. Thus, referential integrity is violated when (1) a foreign key is wrong and, therefore, cannot have a corresponding tuple in the referenced entity or (2) a foreign key is correct, but the referenced entity does not contain the corresponding tuple."(Fürber, 2016)
	Incorrect/outdated reference (Single source): A reference is regarded as outdated or incorrect if the reference is obsolete or between two entities the attribute contains foreign keys referring to wrong tuples in the referenced entity (Fürber, 2016). An incorrect or outdated reference could also occur when a relationship between entities changed over time but not updated in the data source and for example, the geographical address of a

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 101017385

	user could change over time, and a reference link to a weather information database could be outdated or changed into a new link.
	Conditional Missing Values (Single source): A conditional value is defined as an attribute that requires a value only in a certain context, and mainly when the other attributes obtain certain values. For example, the attribute state may only require a value when the attribute country has the value "USA" (Fürber, 2016).
	Mis-fielded values (Single source): "Mis-fielded values are correct values that do not fit the intention of their attribute but to another attribute of the same tuple. For example, the attribute city comprises the value "Germany" which should be located in the attribute country of the same tuple."(Fürber, 2016)
	Heterogeneity of syntaxes (multiple sources): "Attribute values may represent the same real-world entity or state but use different syntactic representations. E.g. there are several different possibilities to represent the current date, for example in the format "dd.mm.yyyy" or in the format "mm/dd/yyyy". Heterogeneity of syntaxes also encompasses the representation of attribute states via cryptic values or codes. In this context, it is also called heterogeneity of representation."(Fürber, 2016)
	Heterogeneity of units of measurement (multiple sources): "The same real-world concept may be represented using different scales. For example, the weight of an object may be represented in one data source using grams, while another data source represents the weight in pounds. Heterogeneity of units of measurement is also known as a data scaling conflict."(Fürber, 2016)
	Data granularity mismatch (multiple sources): "Two or more attributes coming from different sources may refer to the same entity but on different levels of granularity. Data granularity mismatches typically occur when data with different aggregation levels are integrated "(Fürber, 2016). For example, the table "BloodGlucoseLevels" of a user data source may contain the daily average blood glucose levels and another table with detailed blood glucose values for a certain period. Hence, the data cannot be easily compared or joined, since they contain summarized values on different levels of detail. Data granularity mismatches are also known as aggregation or generalization conflicts.
	Default value conflicts (multiple sources): "Different data sources may assign different default values for semantically similar attributes in absence of the real-world information. For example, the attribute LegalAge of data source one may have the default value "18" to indicate adults, while data source two may assign the default value "21" for the same purpose."(Fürber, 2016)
Quality Problems of Data Models	**Outdated conceptual elements (Single source)**: "Conceptual elements, i.e. attributes, tables, relationships, and constraints may become obsolete over time "(Fürber, 2016).
	Missing conceptual elements (Single source): "Sometimes conceptual elements may be missing in the data model, e.g. when a new kind of information becomes relevant that has not been represented in the data model before. Thus, attributes, tables, or other conceptual elements may be missing"(Fürber, 2016). For example, assume the WARIFA risk prediction model requires new parameters that have been discovered recently to have greater health risk than considered before and this requires updating the data model to hold additional attributes.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 101017385

	Misuse of conceptual elements (Single source): "Existing schema elements may sometimes be used to store data values that do not fit the intension of the schema element due to misinterpretation of the semantics of the schema element or inflexibility to extend existing schemata" (Fürber, 2016).
	Overlapping concepts/role conflicts (Single source): "A real-world entity can be part of two or more different real-world concepts at the same time. The concepts may have very different semantics, but due to the membership of the individual to both concepts, they are not disjunctive" (Fürber, 2016). For example, assume a user has both diabetes and skin cancer diagnosed, but the data model design only allows the membership of each user in one of the classes. In many cases, this shows a lack of normalization of the database schema (Fürber, 2016).
	Heterogeneity of integrity constraints (multiple sources): "The constraints on two or more semantically similar attributes can be inconsistent with each other. For example, the attribute age in data source one requires values higher than 18, while the attribute age in data source two requires values higher than 21" (Fürber, 2016).
	Schema isomorphism conflict (multiple sources): "Semantically similar real-world concepts can be represented by a different number of attributes in different data sources" (Fürber, 2016). For example, patient data may be represented in one data source by a table *Patient* with attributes *patient_ID, name, and gender*, while in data source two the same information is represented within a table *Patient* with attributes *patient_ID, name, male and female*.
	Schematic discrepancy (**multiple sources**): "If the schematic differences are not only related to the number of attributes, but the same information is also represented by different schema elements, i.e. data values, attributes, or tables, then we can call this a schematic discrepancy. There are three different types of schematic discrepancies" (Fürber, 2016), i.e. - "data value attribute conflicts occur when the value of an attribute in one database corresponds to an attribute in another database"[1, 2]. - "attribute entity conflicts occur when the same entity is being modelled as an attribute in one database and a relation in another database"(Fürber, 2016). - data value entity conflicts.
Common Linguistic Problems	**Existence of synonyms**: "Two or more values, instances, or names of conceptual elements can be identical in meaning but denoted with different terms" (Fürber, 2016). For example, the attribute chronic condition contains the synonyms values "Diabetes mellitus" and "Diabetes" which represent the same chronic conditions. Synonymous values, instances, and conceptual elements are especially problematic during data integration and aggregation since the synonym relationships must be known to produce precise results.
	Existence of homonyms and polysemes: "Two or more values, instances, or names of conceptual elements can be denoted with the same term but represent a totally or partly different real-world entity (Fürber, 2016). Homonyms may, therefore, easily lead to data quality problems because of misinterpretations. The term "polyseme" is sometimes used interchangeably for homonyms, although it has a slightly different meaning. A polyseme is a word or a sign that has two or more different senses, but the senses are related to each other in opposite to homonyms which can have unrelated meanings" (Fürber, 2016).

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 101017385

	Existence of hypernyms: “A word is a hypernym to another word if it represents a more general meaning than the second one. Hypernymy can be particularly relevant for DQM among pairs of names for tables, attributes, entities, and values. It is then for example difficult to identify the proper semantic relationship in multi-source scenarios. Also, a database manager may map respective conceptual elements with an equivalence relation instead of a proper subtype or type of relation, which can hamper the proper interpretation of the original data at a later point. Data granularity mismatches are frequently caused by the existence of hypernyms” (Fürber, 2016).

4 WARIFA DATA MANAGEMENT PLAN

The tables below summarize the data management plan within WARIFA. It reports on input and output data from the project, without evidencing the data transfers within the project’s WPs. This is because, collected data may come from a variety of data sources, not all of which contain data in a suitable format. Before being used in the development of AI the data need to be prepared and filtered, based on a certain priority and then translated into a communication standard that is compatible with the other modules of the WARIFA architecture. WP3 specific datasets will be used in WP4, WP5 and WP6 for the purpose of developing the AI algorithms described in the DoA.

Dataset Name	WP2 Mapping
Description	Answers to surveys to potential users
Data Type	.doc, .ppt, .xls, other depending on the specific used tool for online survey (CSV, PDF, SPSS) to online surveys
Size	TBD
Language	Romanian/ Norway/Spanish/English
Data About People	Yes
Level of Anonymization	Anonymous
Security Classification	Sensitive data
Collection/Creation Method	Online surveys, focus groups, interviews
Storage	TSD (as described in D1.11)
Transfer	Local/centralized to be decided by consortium
Archiving	5 years after the project end at TSD (as described in D1.11).

Dataset Name	WP3Sensor data
Description	Blood glucose, Physical activity, heart rate, weight, nutritional data, GPS-location, sleep-related data, blood pressure, oxygen saturation,
Data Type	Numerical values + text values + time

This project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No 101017385

Size	Large dataset depending on sample period, i.e. most frequent sample will be every 5. Minute.
Language	Norwegian, Spanish, Romanian, English
Data About People	Yes, personal data
Level of Anonymization	Pseudonymous data (Pseudoanonymization procedures described in D1.11)
Security Classification	Sensitive data
Collection/Creation Method	Collected by sensors automatically, semi-automatically (user trigger collection), and manual.
Storage	TSD (as described in D1.11)
Transfer	Data sent from sensor via app on mobile phone, smartwatch or similar, to TSD. Using secured, wireless communication protocol.
Archiving	For the purpose of WARIFA-implemented functionalities: until project end. 5 years after the project end at TSD (as described in D1.11).

Dataset Name	**WP3 Ubiquitous data**
Description	Demographic data, health-data via questionnaires (alcohol consumption, smoking, medical history, etc,)
Data Type	Numerical values + text values + time
Size	Large dataset depending on sample period and number of individuals
Language	Norwegian, Spanish, Romanian, English
Data About People	Yes, personal data
Level of Anonymization	Pseudonymous Data (Pseudoanonymization procedures described in D1.11).
Security Classification	Sensitive data, need to be high security. GDPR-rules + national rules.
Collection/Creation Method	Collected by questionnaires through apps on mobile phone or PC. Mainly manually.
Storage	TSD (as described in D1.11)
Transfer	Data sent from app on mobile phone, PC or similar, to TSD. Using secured, communication protocol.
Archiving	For the purpose of WARIFA-implemented functionalities: until project end. 5 years after the project end at TSD (as described in D1.11).

Dataset Name	**WP3 Registry data**
Description	Weather data/solar intensity/UV-levels, air pollution, clinical data, etc.
Data Type	Numerical values + text values + time
Size	Large dataset depending on sample period and number of individuals
Language	Norwegian, Spanish, Romanian, English

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 101017385

Data About People	Yes, personal data, and some other kinds of data (weather, pollution, etc.)
Level of Anonymization	For person-sensitive data: Pseudonymous Data - Data with identifiers replaced with artificial identifiers and held separately with technical safeguards. Mostly this will be done in collaboration with the data owners. For the public data: no need for anonymization.
Security Classification	For person-sensitive data: need to be high security. GDPR-rules + national rules. For the public data: no need for security.
Collection/Creation Method	Collected automatically, semi-automatically (user-trigger collection), and manual.
Storage	TSD (as described in D1.11)
Transfer	Data collected from registries, both public and non-public sources, and transferred to TSD. Using secured, communication protocol.
Archiving	For the purpose of WARIFA-implemented functionalities: until project end. 5 years after the project end at TSD (as described in D1.11).

Dataset Name	**WP7 Bayesian belief network**
Description	Causal dependency probabilistic graph
Data Type	Numerical
Size	0.1-1 Mbyte per patient
Language	Numbers+english
Data About People	Yes
Level of Anonymization	Pseudonymous Data (Pseudoanonymization procedures described in D1.11).
Security Classification	Sensitive data
Collection/Creation Method	Output of WP5
Storage	TSD (as described in D1.11)
Transfer	To be decided by consortium
Archiving	For the purpose of WARIFA-implemented functionalities: until project end. 5 years after the project end at TSD (as described in D1.11).

Dataset Name	**WP7 Risk scores**
Description	Probability estimation of patients/subject risk score for specific chronic conditions
Data Type	Numerical
Size	1-100 Kbyte per patient

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 101017385

Language	Numbers+english
Data About People	Yes
Level of Anonymization	Pseudonymous Data (Pseudoanonymization procedures described in D1.11).
Security Classification	Sensitive data
Collection/Creation Method	Computer code execution
Storage	TSD (as described in D1.11)
Transfer	To be decided by consortium
Archiving	For the purpose of WARIFA-implemented functionalities: until project end. 5 years after the project end at TSD (as described in D1.11).

Dataset Name	**WP8 Surveys and interviews**
Description	text and numerical data
Data Type	.doc, .ppt, .xls, other depending on the specific used tool for online survey (CSV, PDF, SPSS)
Size	TBD
Language	Romanian/ Norway/Spanish/English
Data About People	Yes
Level of Anonymization	Anonymous
Security Classification	Sensitive data
Collection/Creation Method	Online survey
Storage	Secure storage in the selected survey tool database, TSD (as described in D1.11) and PNO database (managed in accordance with GDPR regulation, https://www.pnoconsultants.com/privacy-statement/)
Transfer	Download and extraction from the tool to PNO database (managed in accordance with GDPR regulation, https://www.pnoconsultants.com/privacy-statement/).
Archiving	PNO database (managed in accordance with GDPR regulation, https://www.pnoconsultants.com/privacy-statement/), TSD (as described in D1.11) 5 years after the project end at TSD (as described in D1.11).

REFERENCES

BALLOU, D. P. & PAZER, H. L. 1985. Modeling Data and Process Quality in Multi-Input, Multi-Output Information Systems. Management Science, 31, 150-162.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 101017385

CAI, L. & ZHU, Y. 2015. The Challenges of Data Quality and Data Quality Assessment in the Big Data Era. Data Science Journal, 14, 2.

DELONE, W. H. & MCLEAN, E. R. 1992. Information Systems Success: The Quest for the Dependent Variable. Information Systems Research, 3, 60-95.

FRANCISCO, M. M. C., ALVES-SOUZA, S. N., CAMPOS, E. G. L. & DE SOUZA, L. S. 2017. Total Data Quality Management and Total Information Quality Management Applied to Customer Relationship Management. 40-45.

FÜRBER, C. 2016. Data Quality Management with Semantic Technologies, Gabler Verlag.

GOODHUE, D. L. 1995. Understanding User Evaluations of Information Systems. Management Science, 41, 1827-1844.

JARKE, M. & VASSILIOU, Y. Data Warehouse Quality: A Review of the DWQ Project. IQ, 1997. 299-313.

LEE, Y. W., STRONG, D. M., KAHN, B. K. & WANG, R. Y. 2002. AIMQ: a methodology for information quality assessment. Information & Management, 40, 133-146.

MAYDANXHIK, A. 2007. Causes of Data Quality Problems. Data Quality Assessment. Techniques Publications LLC.

PIPINO, L. L., LEE, Y. W. & WANG, R. Y. 2002. Data quality assessment. Communications of the ACM, 45, 211-218.

REDMAN, T. C. 2001. Data quality: the field guide, Digital Press.

STRONG, D. M., LEE, Y. W. & WANG, R. Y. 1997. Data quality in context. Communications of the ACM, 40, 103-110.

WAND, Y. & WANG, R. Y. 1996. Anchoring data quality dimensions in ontological foundations. Communications of the ACM, 39, 86-95.

WANG, R. Y. & STRONG, D. M. 1996. Beyond Accuracy: What Data Quality Means to Data Consumers. Journal of Management Information Systems, 12, 5-33.

ZMUD, R. W. 1978. An Empirical Investigation of the Dimensionality of the Concept of Information. Decision Sciences, 9, 187-195.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 101017385