CrowdHEALTH

Collective Wisdom Driving Public Health Policies

Del. no. – D5.15 Multimodal Forecasting and Causal Techniques: Design and Open Specification

Project Deliverable

This project has received funding from the European Union's Horizon 2020 Programme (H2020-SC1-2016-CNECT) under Grant Agreement No. 727560

D5.15 Multimodal Forecasting and Causal Techniques: Design and Open Specification

Work Package:		WP5
Due Date:		31/10/2017
Submission Date:		06/11/2017
Start Date of Project:		01/03/2017
Duration of Project:		36 Months
Partner Responsible of Deliverable:		Karolinska Institutet
Version:		1.1
Status:	☒ Final ☐ Draft ☐ Ready for internal Review ☐ Task Leader Accepted ☐ WP leader accepted ☒ Project Coordinator accepted	
Author name(s):	Sokratis Nifakos (KI)	Mitja Lustrek (JSI) Anton Gradisek (JSI) Gregor Jurak(ULJ) Maroje Soric(ULJ) Bojan Leskošek(ULJ) Thanos Kosmidis(CRA) Christos Panagopoulos (BIO)
Reviewer(s):	Jan Jannsen (DFKI) Andreas Menychtas (BIO)	
Nature:	☒ R – Report ☐ D – Demonstrator	
Dissemination level:	☒ PU – Public ☐ CO – Confidential ☐ RE – Restricted	

REVISION HISTORY			
Version	Date	Author(s)	Changes made
0.1	01/07/2017	Sokratis Nifakos	TOC to share with the participants.
0.2	23/08/2017	Sokratis Nifakos	Basic framework
0.3	14/09/2017	Sokratis Nifakos	Executive summary and contents
0.4	21/09/2017	Sokratis Nifakos	Causal analysis introduction and description
0.5	22/09/2017	Sokratis Nifakos	Causal analysis in public health sciences
0.6	22/09/2017	Sokratis Nifakos	Parts 3.4.5.6
0.7	09/10/2017	Sokratis Nifakos	Parts 4.5.6 updated
0.8	13/10/2017	Sokratis Nifakos	Parts 3.1 till 3.2
0.9	16/10/2017	Sokratis Nifakos	Integrated information from CRA and ULJ
1.0	30/10/2017	Sokratis Nifakos	Revised document after the peer review
1.1	06/11/2017	ATOS	Quality review and submission to EC

List of acronyms

ANM	Additive Noise Methods
ARMAX	Autoregressive moving average with exogenous inputs
BMI	Body Mass Index
BN	Bayesian Network
COPD	Chronic Obstructive Pulmonary Disease
CRA	CareAcross
CVD	Cardiovascular disease
FVC	Forced Vital Capacity
HULAFE	Hospital LA FE
IGCI	Information Geometric Causal Interference
IP	Inverse probability
IPF	International Powerlifting Federation
NN	Neural Network
SVM	Support Vector Machine

Contents

Table of figures / tables

1. Executive Summary

This document is the causal analysis framework of the health policies toolkit and it will be used as the base for the development of the software prototype that applies to the health analytics layer in CrowdHEALTH architecture (D5.15). This framework is focusing on the analysis of actions and events in the Use Case scenarios aiming to estimate the applicability and the effectiveness of the current health policies referring to the specific case. Moreover, based on the information collected, the causal analysis framework will provide information to the forecasting model in order to produce future recommendations to the policy makers.

In particular the following topics will be specified here:

1. An introduction to causal analysis for public health policies.
2. The causal analysis and forecasting methods.
3. The Causal Analysis framework design.
4. The related information from the Use Case Partners.

2. Introduction

The questions that motivate most studies in the health, public health, social and behavioural sciences are not associational but causal in nature. For example, what is the efficacy of a given drug in a given population? What fraction of past clinical mistakes could have been avoided by a given policy? What was the cause of death of a given individual, in a specific incident? These are causal questions because they require some knowledge of the data-generating process and they cannot be computed from the data alone, nor from the distributions that govern the data. [1] Conceptual frameworks and algorithmic tools needed for tackling such problems.

Why causal and not statistical analysis? The aim of standard statistical analysis, typified by regression, estimation, and hypothesis testing techniques, is to assess parameters of a distribution from samples drawn of that distribution. [1] With the assistance of such parameters, one can infer associations among variables, estimate beliefs or probabilities of past and future events, as well as update those probabilities in light of new evidence or new measurements. As long as experimental conditions remain the same, standard statistical analysis is the best method to manage well these tasks. [2]

Causal method goes deeper in the analysis and it's aiming to infer not only beliefs or probabilities under static conditions, but also the dynamics of beliefs under changing conditions, for example, changes induced by treatments or external interventions. There is nothing in the joint distribution of symptoms and diseases to tell us that curing the former would or would not cure the latter. More generally, there is nothing in a distribution function to tell us how that distribution would differ if external conditions were to change—say from observational to experimental setup—because the laws of probability theory do not dictate how one property of a distribution ought to change when another property is modified. This information must be provided by causal assumptions which identify relationships that remain invariant when external conditions change. [2]

3. Causal Analysis and Forecasting design and open specification.

3.1 Descriptive structure and causality

There are different approaches for analysing clinical and public health data depending on different factors such as whether the investigator intervenes in process or not. Thus, in the experimental approach the investigator intervened in the experiment with in the sense that it controls the distribution of individuals within groups (usually assigns randomly treatments for each person involved). [3] In contrast, in the observation studies the researcher did not intervenes, simply observes and records. In this case the main purpose is the investigation of the causes of a disease and constructing public health policies data models.

Another separation of medical studies is made depending on whether its purpose is simply to describe a population or if we want to understand relationships between features and to identify the causes that cause a phenomenon, usually a disease. [4] Descriptive studies are limited to simple descriptors indicators or demographic tables and a graphical representation of the data while the causes systematically apply statistical tests of assumptions based on causality theories or scenarios. Additionally, it is common to use complex statistical models.

3.1.2 Descriptive structure

Descriptive approach simply aims to describe a population at a specific place and time. For this reason, we use the usual sampling patterns when we have large populations or censuses when the population is limited (usually in demographic studies). [5] These studies are also called cross-sectional. When the purpose is limited to estimating the prevalence (proportion of people with the disease in the population) they are called demographic or prevalence. One of the biggest advantage is that they can present a general picture of the relationships between diseases something which cannot be done with in-hospital investigations. Moreover, some other advantages of this approach are: the simplicity, time efficient and low cost. [6] On the other hand, disadvantages include the fact that cannot take into account the factor of time (hence we cannot indicate the time sequence of relationships between variables), we cannot also distinguish the likelihood of occurrence and duration of the disease and finally rare diseases are not eligible for studies since the size of the sample is usually inadequate (so they are limited to more common diseases). [6] All of the use cases in CrowdHEALTH platform focusing on common diseases in order to avoid these limitations.

3.1.3 Causal analysis in public health domain

Causal analysis is the backbone of modern epidemiology. It refers to studies that aimed to control an epidemiological case and also to investigate the causality of a disease. In CrowdHEALTH analytics layer this research approach will be based on sophisticated statistical methods starting from routine statistical tests (for example t-test or Wilcoxon test) up to complex models such as hierarchically generalized linear models and multivariate analysis models.

One of the aims of CrowdHEALTH project is to include both observational data (such as prospective and retrospective data) as well as interventional data for instance clinical trials. The main purpose of this observational data is to identify risk factors which increase the incidence of a disease and how this information could improve the diagnosis and prevention strategies, while in clinical trials the purposed is to identify new drugs that limit their symptoms of a disease or lead to complete healing (treatment). The development of holistic health records will enable to include both data in one dataset.

3.1.3 Variables' relationship and causality

One of the biggest problems in statistical science is the fact that a statistically significant association or association between two variables does not automatically imply on a real inductive or causal relationship. This is because of the complexity of relationships that exists in phenomena we are trying to investigate. [7] Therefore, in occurrence of a phenomenon (eg disease) is caused by many factors and for this reason the risk-factor relationship can be influenced by third known confounders. Further information will be provided later in this document.

Confounder or confounder variables called the variables that distort a relationship between two variables. For instance in CrowdHEALTH project based on the different use cases we usually mean the disease and a risk factor. For instance in Karolinska's Institutet use case scenario which is focusing on cardiovascular diseases, we suppose that we want to compare mortality in two groups with different exposure to a potential risk factor (e.g. smoking). If the first group contains elderly people and we expect greater mortality in this group that would not be due to the fact that they are smokers but simply because of the fact that they are old, then age is a confounder of the mortality relationship of smoking. A confounding agent may display as statistically significant relationships that do not actually exist or cover existing ones. The control and adhesion of confounders is done by appropriate study design or by specific statistical analysis based on the use case scenarios.

Even if confounding factors are eliminated, the identification of real relationships causality is difficult and goes beyond the simple calculation of statistical indicators. For that purpose in this task we have set specific criteria that we can use to identify and verify whether the statistically significant correlations are also relationship based causation. [8] These criteria are as follows:

1. Consistency;
2. Strength;
3. Specificity;
4. Temporality;
5. Coherency.

When a relationship meets all the above criteria, then we have a strong (if not absolute) indication that this is a causal relationship.

Next we will describe further each of criteria.

3.1.4 Evaluation criteria for causality relationships

1. Consistency of a relationship

A relationship is called consistent if it repeatedly appeared in data that have been generated with different design and for different populations. For instance in Care Across use case scenario a consistent connection would be the positive relationship between smoking - lung cancer that it has been confirmed in a large number of studies during the previous years.

2. Strength of a relationship

A relationship is called strong when the effect of the presence risk factor is high. The effect is usually measured by the proportional increase of death possibility or the occurrence of the disease. In this case, it is important the existence of the dose response effect which is the proportional variation of the variable response (e.g., occurrence of the disease) and the prescript dosage of a drug or the size of exposure to a risk factor. Considering again the Care Across use case for example the probability of lung cancer is increased in smokers and it increases accordingly to the number of cigarettes that each person smokes.

3. Specificity of a relationship

A risk-disease factor relationship is called if a presence of the risk factor leads to a high probability of developing the disease while its absence leads to a high probability of avoiding the disease. If one agent is marked as high weighted factor, (it's the major cause of a disease) then it is very likely that the relationship is causative. In practice, however, because of the complexity of relationships, specified relationships are rare. Moreover, the causation of the disease may increase significantly with the presence of the risk factor, but not so much as to characterize it specific. The existence of a statistical causal relationship does not imply that it will be verified in all cases. An example, if someone smokes does not mean they will show the disease surely (if this was the case, we would not be talking about a statistical relationship other than a physical relationship).

4. Temporality

In many cases the statistical analysis does not take into account the time sequence or sequence in which some contingencies appear and also the variables associated with them. This is usually done by using common sense. For this reason, the response variable in the statistical models should be followed in time from the explanatory one (something that is not always obvious). In conclusion it is necessary for the risk factor to occur before the onset of the disease. In some cases this is easy to judge for instance in the smoking and cancer lung scenario. But what about cases such as passive smoking for which it cannot be is easily

detected the exposure time to this risk? Defining temporality is also difficult in cases where the disease is diagnosed much later than its actual appearance.

5. Coherency

In order for a relationship to have a logical sequence it should not contradict to proven truths of physical and biological sciences for instance Physics, Chemistry, Medicine, Genetics and Biology. When a statistically significant relationship contradicts an already recognized theory, then there should be another scientific research that will support and justify its existence.

3.1.5 Secondary notional relationship between two variables

Given the existent of a real causal point of a relationship between two variables then we expect to see a statistically significant relationship between them. The opposite is not always true, meaning that we can have a significant statistical relationship which is not due to a realistically existent one. [9] One simple example could be presented from the CareAcross use case where we suppose that we want to look the association between smoking and lung cancer. Smoking causes an increased incidence of cancer but also yellowing fingers in case of heavy smokers and smoking time period. If we isolate the variable that indicates the yellowed fingers (ignoring the smoking that causes them) and the disease, then it is most likely to have a statistically significant relationship, that is, people with yellowed fingers have greater possibility of cancer or that cancer causes the appearance of yellowed fingers). In fact, the relationship is fictitious and it's just because those who have yellowed fingers smoke a lot more in relation to the rest so obviously smoking is what causes it.

Therefore, in a notional relationship between two independent variables B and C, these variables present to have a statistically significant dependence on their relationship to the variable A which is the cause of both. If the distribution of variable C will change, it does not necessary mean that the distribution of variable B will change as it would be changed if the relation was really causative. [9]

In this task it is very crucial to identify through the different use case scenarios, the notional relationships as it is tightly connected with the risk stratification module in CrowdHEALTH platform where the identification of risk factor in a population based sample is the main aspect.

3.1.6 Direct and indirect causality relationships

It is important to present the direct and indirect causality relationships and the importance of them. We will use the HULAFE use case scenario in order to present an example of these relationships. If variable A affects the occurrence of the variable D and that in turn variable B then we say that variables A and B are linked to each other by an indirect causation relationship. Traditionally in public health and medical sciences, the factors associated with a disease with direct causation are considered to be more dangerous and important, while indirect risk factors are considered to be subsidiary or predisposition. [10]

One example from our use cases is from HULAFE which deals with obesity patients. Obesity causes among others, hypertension and diabetes and these in turn coronary heart disease. Thus, the relationship between coronary artery disease and obesity is indirect, meaning that if a patient has been diagnosed with obesity but does not show other heart disease symptoms then he will not registered as a high risk patient for coronary heart disease. However, and because obesity is strongly associated with other diseases it is very important to take into account any possible indirect relationships in CrowdHEALTH use cases.

3.2 Multimodal forecasting

3.2.1 Prospective data analysis

The multimodal forecasting approach in this task it has the meaning of analysing group of patients or population based samples that we observe over time. The key element is through this time based observation we are able to explore their progression towards some diseases and potential risk factors. [11] In most cases in public health studies, we are repeating the measurements in order to end up to the right conclusions but within the CrowdHEALTH platform we aim to simulate these metrics in order to forecast diseases development or the potential risk factors.

The basic characteristics of the multimodal forecasting framework are the avoidance of errors due to the record keeping or memory recalling, the importance of time and time sequence of relationships and the feasibility to estimate the variance that caused from the individuals. [12]

3.2.2 Historical health records and cohort approach

Most of the data from our use case partners in CrowdHEALTH project is historical. For this manner, the forecasting analysis will be focused on the patients groups of interest (cardiovascular, obesity, cancer etc.) and analyse their data as objectively “clean” historical records. Although it seems essentially retrospective analysis, it will be used in this task prospectively in the sense that the data obtained from the use case scenarios is clean and they don’t include time based errors so it can be used in a simulated test scenario. It has to be mentioned though that the risk of errors remains high and it is related to the accuracy and the validity of the data.

3.2.3 Binary approach in public health cases

In most of the clinical and epidemiological studies the main response variable is binary, meaning that the disease is present or not. In addition, the most common risk factors for a disease development are also classified, such as gender. [13] Thus, in this task we will define indicators for measuring the risk of disease development when the classifiers are binary. In the simplest case, we are interested in forecasting the risk of a disease to a person exposed to the risk factor with the corresponding risk of an individual who is not exposed to the same risk factor.

In a population based forecasting risk factor analysis we are interested in the main value of the compared groups in relation to some classified variables that we are interested in comparing the reserved distributions or probabilities. This practically translates into comparing the percentages (probabilities or ratios) for each different exposure group to a risk factor.

4. Classifiers

In this section we will describe a data driven ensemble classifier which will be used for population and individual based disease analysis based on causal variables and health records. We have chosen to use a class-wise classification as a pre-processing step in order to obtain an efficient ensemble classifier. The logic we follow is to combine a variety of classifiers, either different type of classifiers before we make a final classification decision. In that way we enable the distribution of different characteristics of the samples to be handled by the appropriate classifiers that serve their particular needs and provide an extra feature of bias trade-off.

First we will pre-process the datasets into more homogeneous cluster groups and in continuous we will apply the ensemble classifier in order to predict the categories of patients. With this method we plan to produce more accurate and easily interpreted classified results.

Our ensemble classifier architecture consists of four phases:

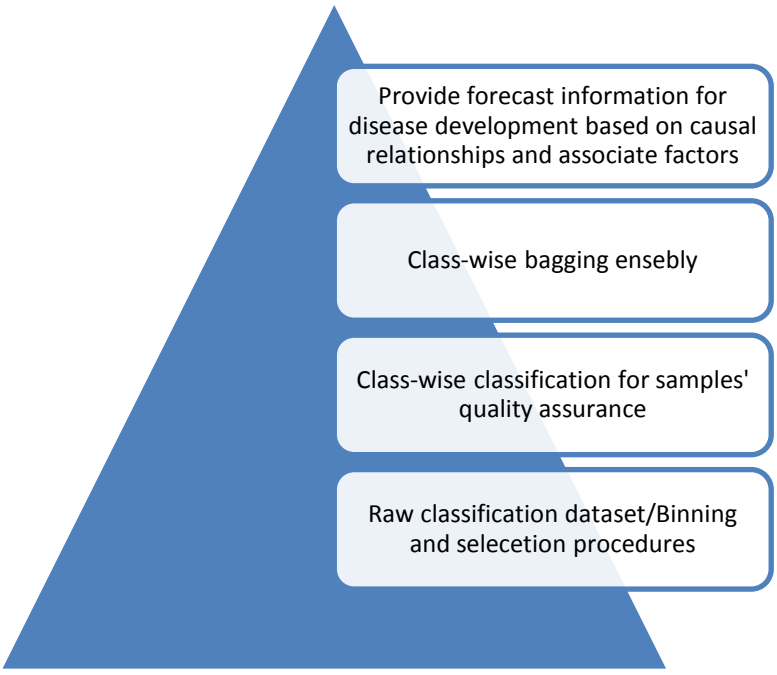


Figure 1: Classifier Architecture

After the pre-process of the datasets and the raw classification, the class-wise classification will fusing three types of classifiers. Neural Network (NN), Bayesian Network (BN) and Support vector machine (SVM).

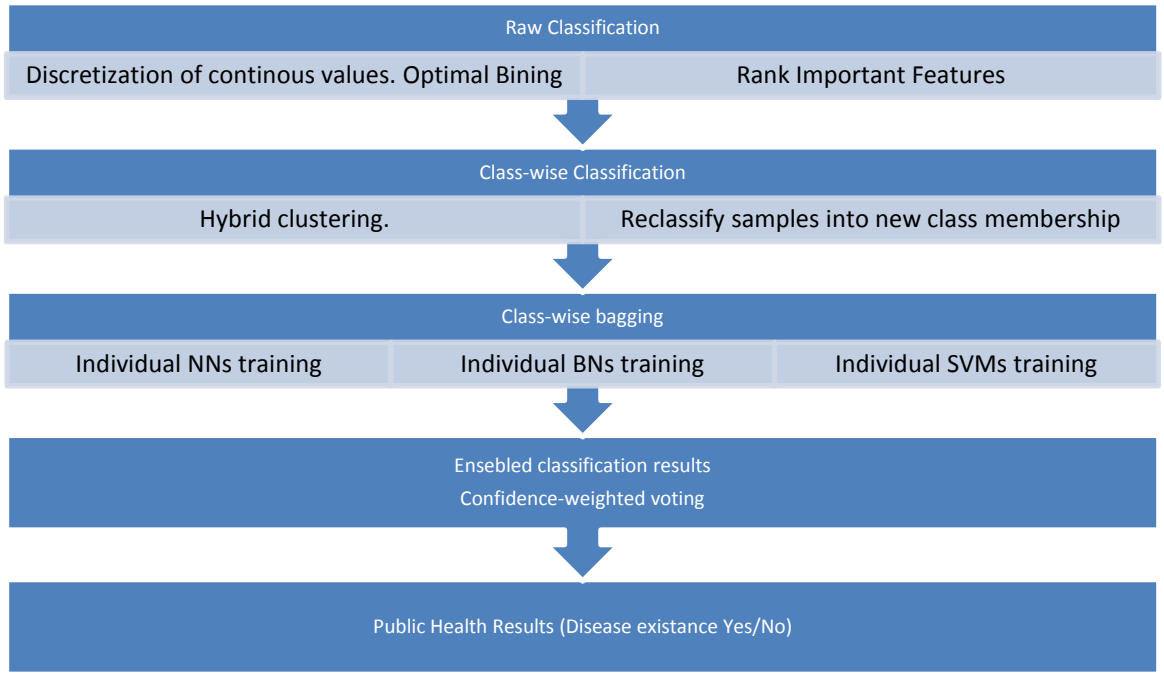


Figure 2: Data Flow

The ensemble classifier is valuable due to its ability outperform the best individual classifier's performance. Ensemble classifiers use the idea of combining information from multiple sources to reduce the variance of individual estimation errors and improve the overall classification results. Importantly, the key to successfully building an ensemble classifier system is to construct appropriate input training sets and maintain a good balance between diversity and accuracy among individual classifiers in an ensemble. For that reason, the ensemble classifier will use as an input the holistic health record repository in CrowdHEALTH platform.

Bayesian Network (BN)

Bayesian network is a probabilistic graphical model that represents a set of random variables and their conditional dependencies via a directed acyclic graph. It can be used to build models from data and/or expert opinion. Each node in the network is associated with a probability function that takes, as input, a particular set of values for the node's parent variables, and gives (as output) the probability (or probability distribution, if applicable) of the variable represented by the node. BN are widely used in the field of disease prediction. Within the

framework of CrowdHEALTH, BN can be used to represent probabilistic relationships between diseases and symptoms or risk factors.

Support vector machine (SVM)

As opposed to BN, SVM are non-probabilistic models. They rely on supervised learning on a labelled dataset. SVM builds a binary classification model that assigns new examples to one of the defined classes. In graphic representation, a SVM model can be seen as a hyper-plane in a multi-dimensional space, separating the examples into two classes divided by a clear gap, as wide as possible.

Neural Network (NN)

Neural networks (actually artificial neural networks) are inspired by the biological networks of neurons that are found in brain. A network consists of artificial neurons, each of which processes the signal (input) and then passes it to the connected downstream neuron via a connection (synapse). Neurons are typically organized in layers where each layer performs different kinds of transformations on the input. NN have been recently experiencing a renaissance due to the development of the computer hardware optimized for parallel processing of data, originally used for graphic cards. NN have proven strong for tasks such as image recognition, machine translation, speech recognition, and others. In 2015, a computer program Alpha Go that is based on NN beat a professional human player in the board game Go. In medicine, the potential of NN is especially strong in image processing for medical diagnostics. The downside of NN models is that they are not comprehensible to outside observer, as opposed to, for example, decision trees, where it is clear what decision is taken at each node.

4.1 Bayesian learning and Markov blanket

One of the main aims of causal analysis and forecasting module in CrowdHEALTH is to discover hidden patterns in the datasets so that policy makers will better understand the different characteristics and causal associations for chronic diseases and develop new public health policy strategies.

The BN model is a powerful knowledge representation and reasoning algorithm under conditions of uncertainty. The major advantage of BNs over many other types of predictive models, such as neural networks, is that unlike ‘‘black box” approaches, the BN structure represents the inter-relationships among the data set features.

BNs offer good generalization with limited training samples and easy maintenance when adding new features or new training samples which are very important in this project taking into account the data availability from the included stakeholders. Based on a general BN classifier, we can get a set of features that are in the Markov blanket of the target feature, and

features provided by the Markov blanket are sufficient for perfectly estimating the distribution and for classifying the target feature classes.

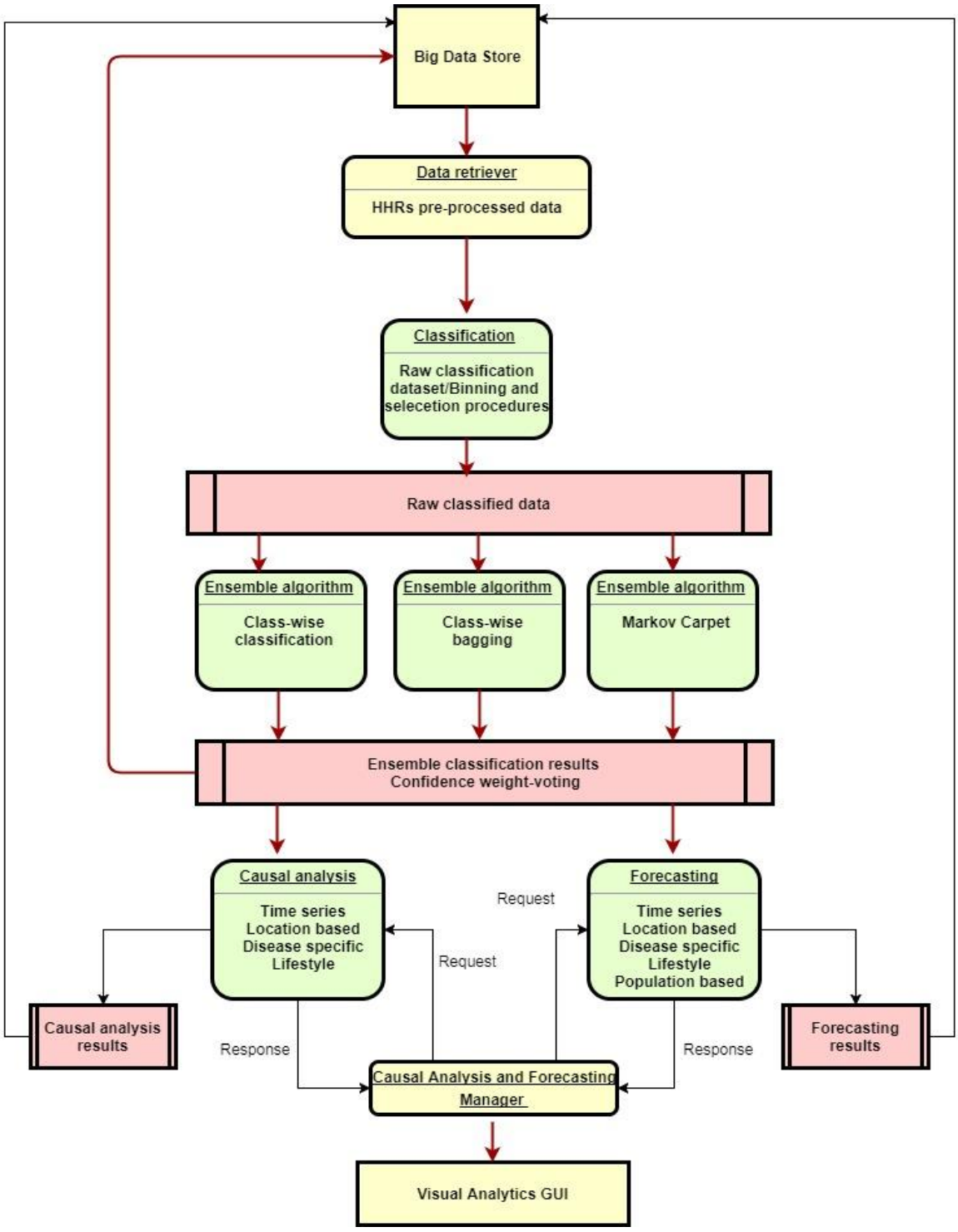


Figure 3: Causal Analysis and Forecasting Architecture

5. Causal Analysis and Forecasting Types

5.1 Overview

As discussed above, there is a prominent difference between correlation and causality. Two variables may be correlated but not causally related. Here, we outline some definitions and methods to determine whether we can talk about causal relationship and how to evaluate it.

There are two principles that define the causality relationship:

- The cause happens prior to its effect.
- The cause has unique information about the future values of its effect.

In this section, we look at three different aspects of causality and at the end outline the plan for the analysis of use cases.

5.2 True Causality

Given joint observations of variables (X, Y), we want to determine which variable is the cause and which one is the effect only by observing the data. If the effect Y is caused by X, we write X -> Y (and vice versa). There are two families of methods that can solve our problem. The first family are **Additive Noise Methods (ANM)** and the second are **Information Geometric Causal Interference (IGCI)** [14].

If we want to use ANM to prove X -> Y, we want to find the relation between variables X and Y, so that Y is equal to some function of X added by the noise E. The E is a variable (usually we take a Gaussian distribution) that is independent of X. Therefore, if we can find such a model to compute Y from X, and cannot find this type of model in that computes X from Y, we may assume X –> Y.

On the other hand, IGCI predicts the direction of causality without bothering about the noise but only focusing on cross distribution of (X, Y). The idea behind this method is that if Y and X are related by some function and if X -> Y, then the distribution of X and the conditional distribution of Y|X contain no information of each other, but distribution of Y and X|Y are somehow related.

5.3 Disentangling causes from confounders

Assuming that we know the direction of causality (we can distinguish between cause X and effect Y) we can in next step evaluate it. As it is plausible that Y is affected by more than just one cause, we want to estimate exactly how Y is related to observing cause X [15]. In this paragraph we will name the cause of our interest treatment and other features that also affect Y covariates (for people, covariates could be: sex, race, education, age...). We want to set up a model that estimates the value Y by knowing the value of X (by example, a linear curve: $Y = c_0 + c_1 X$). However, because this model does not contain all information about the causes of

Y, we have to gather covariates that we think are sufficient for estimating Y. Then, we have to use one of the statistical methods to remove all arrows that point from covariates to Y. A covariate may be of direct interest or it may be a confounder, a variable that influences both the dependent variable and independent variable causing a spurious association.

The first method is **Inverse probability (IP) weighting**, which creates a pseudo-population from our data, in which all effects from covariates are removed. Therefore, on this »new dataset« we can test our causality assumption model X -> Y. This method can also be modified to observe different cause of X to Y between different groups (stratums) in our population (by example: between males and females).

Another method is called **standardization (g – formula)**, which computes the estimate of Y in cause treatment according to sizes of groups (stratums), that are defined by covariates.

5.4 Causality in time series

Time series Y is a (numerical) variable that changes through time. We want to predict the value in the future using known values from the past. This is highly relevant in most of CrowdHEALTH use cases as we would like determine connections between various measurable parameters in present to possible health-related problems in future.

If we have data for time series X and want to determine whether X is also relevant for forecasting Y, we can use the **Granger Causality test**. In other words, we are interested whether the lagged X values provide statistically significant information about future values of Y. We have to choose the lag between the data we want to predict and the data we are using for forecasting and the length of the time interval from which the data are taken. The number of lags can be chosen various information criteria.

First we try to fit only data from Y to a model; second we extend the model and use both data from Y and X for fitting. If the second step improves our prediction about the outcome of Y, then it is wise to use past data from both time series for forecasting Y.

While using the Granger Causality test, one has to be careful not to obtain spurious causalities. Namely, the Granger-causality tests are designed to handle pairs of variables, and may therefore produce misleading results when the true relationship involves three or more variables which have not been initially considered.

5.5 Health forecasting

There are two main groups of quantitative forecasting models:

- Time series models;
- Causal models.

Time series models try to determine the model that explains the historical data and allows for extrapolation until a certain point in future. Some of the approaches that can be applied here include trend curve analysis, moving average or exponential smoothing.

Causal models, on the other hand, search for causal relationship between variables and are preferred when we have some background knowledge on the topic. An example of a model used here is **multiple linear regression**. It uses the causal relationship between the predictive variable and the set of causal factors (i.e. predictors). Another model is **Autoregressive moving average with exogenous inputs (ARMAX)** which uses only the last portion of the time series for predictions.

Looking at the SLOfit use case as the representative example, we are interested in particular in the relationship between the health-related parameters and potential health complications in adulthood. The indicators that we will focus on include obesity (both the BMI and the peripheral fat measurements), cardiorespiratory fitness (evaluated from the 600 m run performance, which is also used to determine the VO2max values), and muscular fitness, which is determined from sit-ups and bent arm-hang. The plan is to use the data obtained up to a specific point in time to extrapolate the values towards the value at the age of 18 (where the measurements end) and further use this knowledge for forecasting the health risks in adulthood. As the data is limited, an extensive combination with relevant data from epidemiological studies will be required. The plan is in more detail explained in Section 6.

6. Data processing

6.1 Data aggregation and accuracy of public health forecasting

When engaging in any analysis, even if not involving causality or forecasting, it is inevitable that some level of aggregation will be performed. This is because as sample sizes increase, the importance shifts from the individual to the population – and this is a shift in the right direction because the objective is to apply the learnings in a broader scope, and not just the people already involved.

The various parameters and levels of aggregation are an important part of the process. Firstly, they are dependent upon the datasets captured. Furthermore, both the data inputs, the parameters and the aggregation levels are inevitably subject to human perception bias: we main implicit assumptions on what could feature be the determinants and drivers of forecasting or causality analysis. Finally, parameters and aggregation levels are to be devised in such a way as to make coherent sense in a public policy setting. (For example, it is unreasonable to aggregate patients with a specific diagnosis and those with a set of treatments incompatible with that diagnosis.)

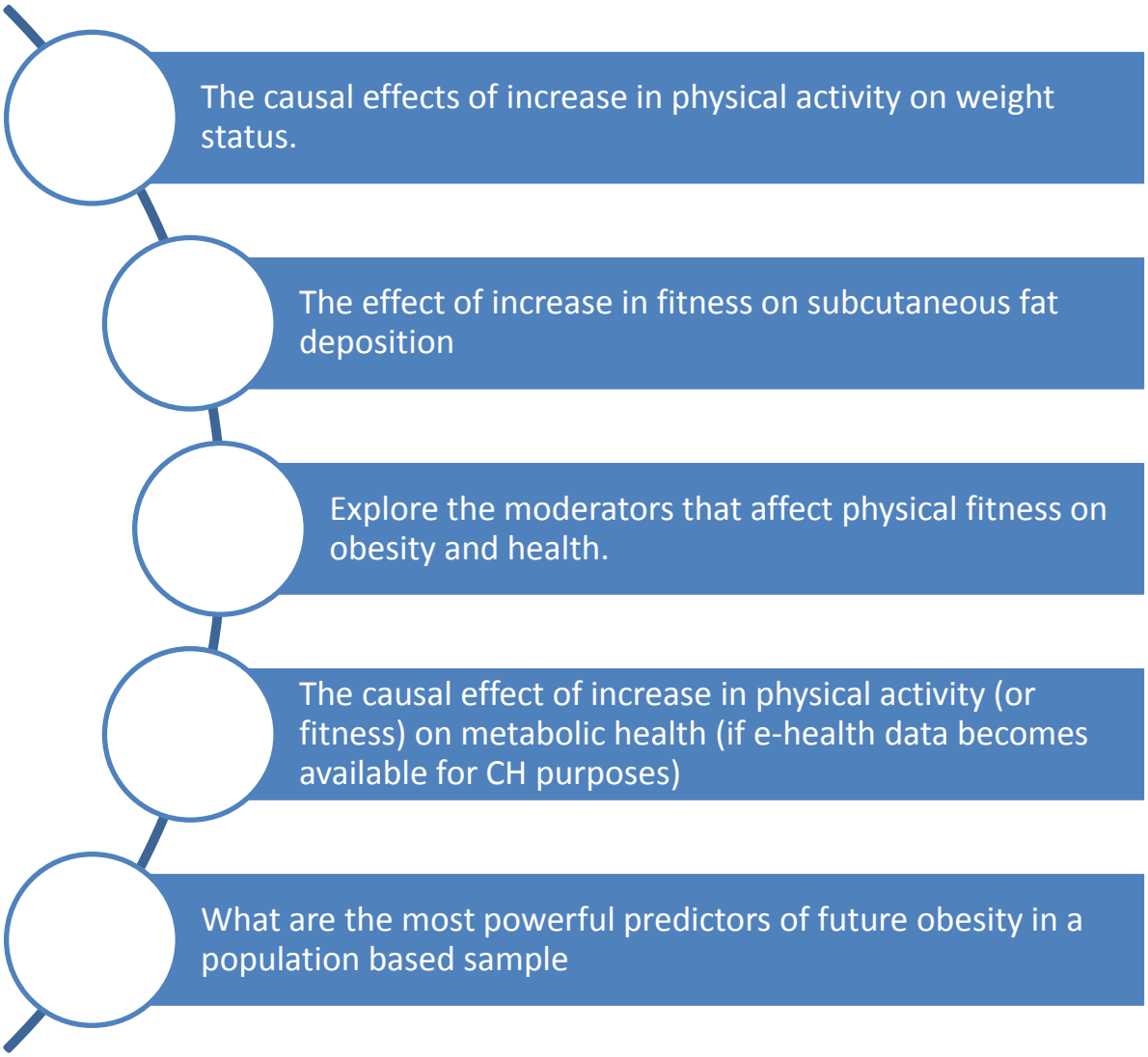
The accuracy of any findings and projections into a generalised audience largely depends on the size of the study. As indicated in the project DoA, the CRA use case will involve 1000 cancer patients. We have designed the cohort to consist of breast cancer patients in order to be able to draw learnings in a consistent way, to the extent possible of course.

Moreover, it is important to note that any patient-reported datasets may suffer from selection bias: a patient who is able and willing to use technology and enter their own data is not necessarily representative of the entire population of patients. Therefore, the accuracy of any finding will need to be re-examined at length for the purposes of accurate public health forecasting.

7. Use Case Causal Analysis and Forecasting

7.1 Causal analysis and forecasting for fitness and obesity

Within ULJ use case it is planned to disentangle the complex causal relationships of obesity and fitness. Specifically, during the first phase of the project the use case will rely on data about fitness and weight status from the existing SLOfit database aiming to identify causality factors and the nature of their relationship as follows:



7.2 Data available and limitations

At the end of year 1, causal analyses will be performed on a dataset containing 8 000 subjects. Each individual has between 1 and 14 data points of weight status and fitness. This longitudinal dataset will provide the opportunity to ascertain temporal relations and rule out reverse causality.

Around month 22 we plan to repeat all analyses on a dataset containing 220 000 subjects between 1 and 14 data points of weight status and fitness. Around this time we expect to have completed the pilot study and have at our disposal a dataset with added lifestyle data on physical activity, sedentary behavior and sleep.

Although we plan to assess some confounders, the main limitation in this use case will be residual confounding that is expected as no data on genetics, nutrition or energy intake is available. In addition, we will need to adjust for measurement bias that is moderate in fitness assessments and large in physical activity assessment.

7.3 Causal analysis and forecasting for chronic diseases

The ability of healthcare institutions to utilize predictions about a patient's status to make appropriate interventions is becoming increasingly important, especially when dealing with complex chronic diseases, as selection of optimal therapy may require integration of multiple conditions and factors that are considered in isolation under current approaches to care. Furthermore, early identification and proper management of risk factors associated with chronic diseases is essential for developing appropriate health policies to intervene at different levels.

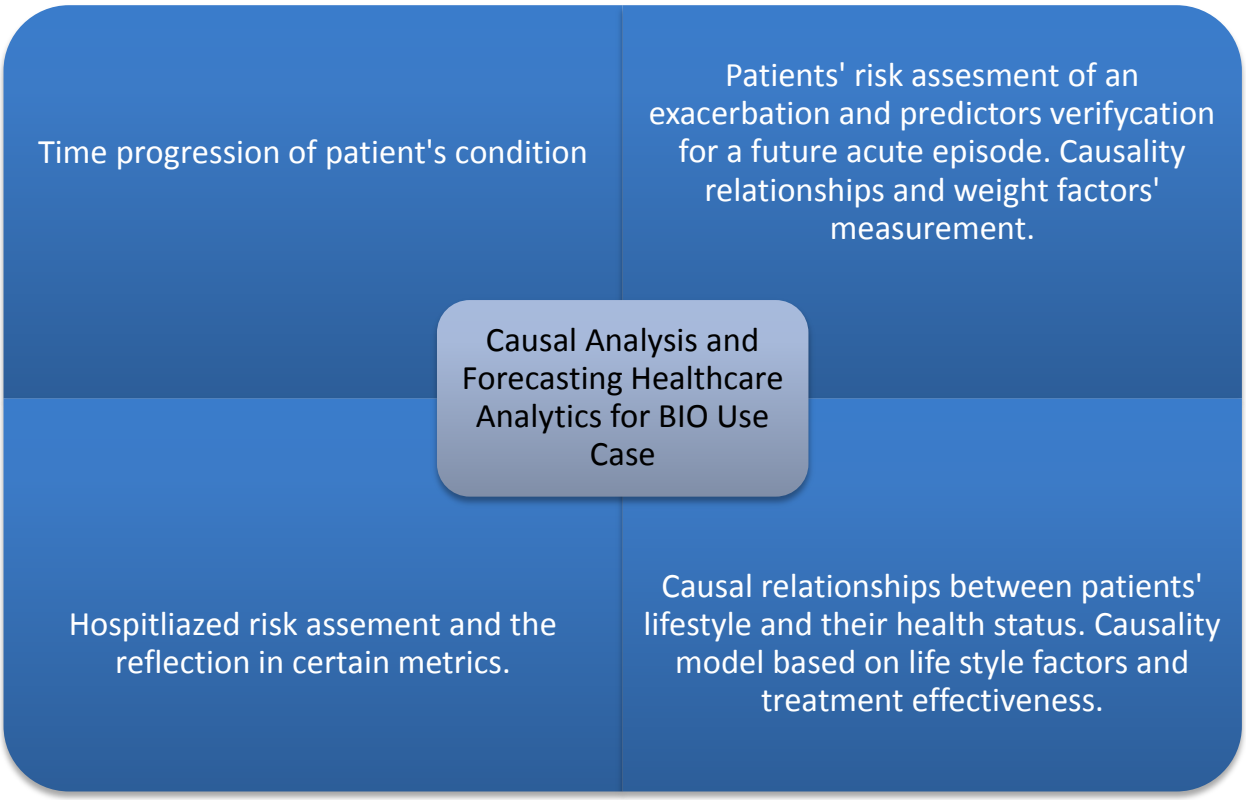
Clinicians treating chronic patients are constantly trying to monitor disease progression, in an attempt to predict exacerbations and acute episodes, hospitalization and mortality. However, predicting future disease trajectory is an extremely challenging problem. One difficulty is the many underlying sources of variability that can drive the different potential manifestations of the disease. Another challenge is the fact that observations are in most cases irregularly sampled, asynchronous, and episodic, precluding the use of many time series methods developed for data regularly sampled at discrete time intervals.

While there is no single approach to tracking and predicting progression for all chronic conditions, monitoring any chronic disease involves measuring certain vital signs and significant – albeit relative to the specific condition – changes in patterns or levels of these measurements are associated with outcomes. For example, decline of 10% in FVC or 3% in SpO2 measurements are considered important indicators of deterioration of the health status of an IPF patient.

Other common parameters that are considered important in most chronic diseases are BMI, physical activity levels and patient adherence. These parameters provide information on patient behaviour and lifestyle and can therefore be used to study the causal relationship

between a patient's ability for self-management and outcomes on their health. In addition, BMI is also used to calculate a variety of indexes used for chronic disease monitoring.

Ultimately, within this use case, the aim of causal analysis and forecasting, relying on data with respect to the aforementioned parameters is to provide insights into the following questions:



7.4 Data available and limitations

There are approximately 100 patients enrolled in BioAssist's pilot sites. These patients have already been performing daily measurements of at least two types of biosignals for almost a year and new data are continuously collected. In addition, information on patient adherence is automatically logged by the BioAssist platform and we have recently started collecting physical activity data. The available datasets are described in detail in deliverable D1.2.

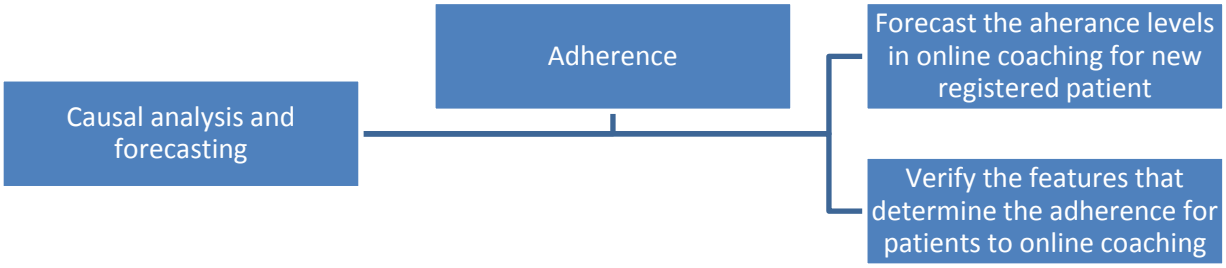
Most of the patients enrolled in BioAssist's pilots suffer from respiratory conditions (IPF or COPD). As such, at the end of year 1, causal analysis and forecasting will focus on answering questions that are relevant to the specific chronic diseases. The results of this effort will act as a starting point to expand the approach to other types of conditions, such as CVD, for which more data will be available for the end of year 2.

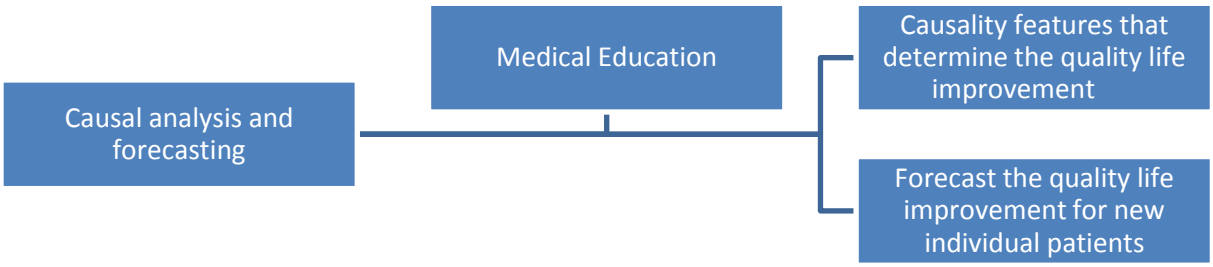
7.5 Causal analysis and forecasting for cancer

From the policy making perspective the CRA scenario is aiming to address the following health policies questions:

1. Can patients adhere to online coaching?
2. Does medical education help patients improve their quality of life?

Consequently, causal analysis and forecasting can derive four very important learnings and can be distributed in two categories, the adherence and the medical education:

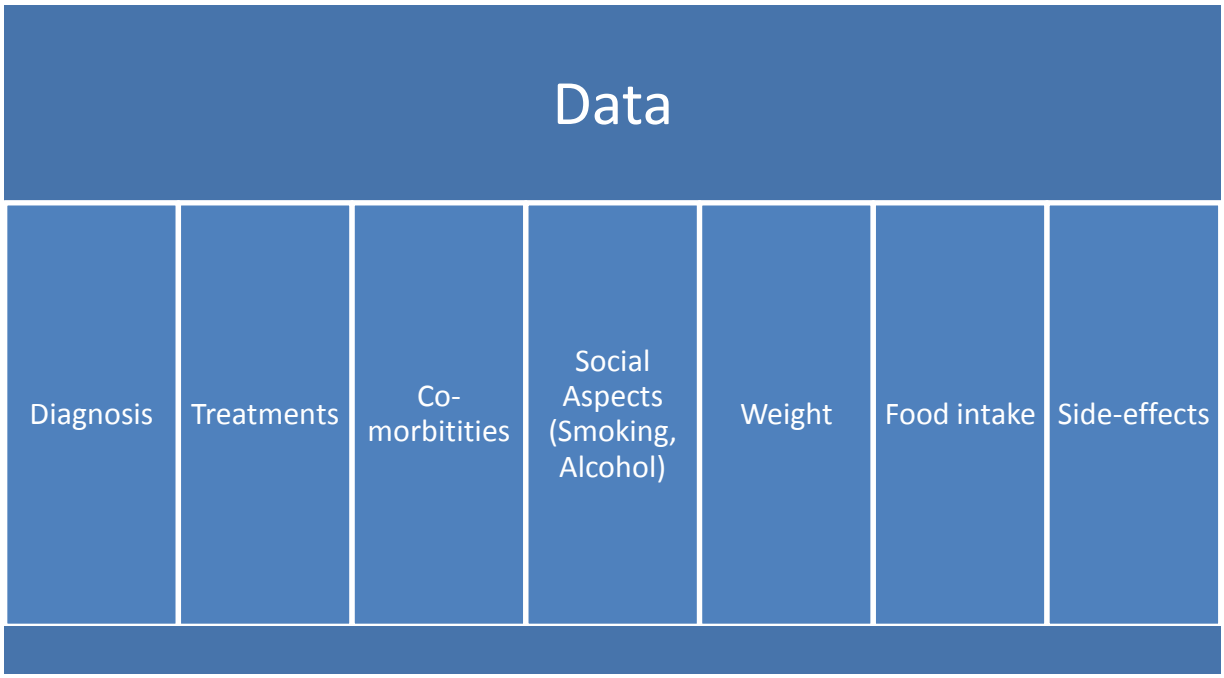




While we do not expect the end results to be complete or entirely deterministic, they will be able to shine a light onto the patient characteristics that can be in focus when devising such programmes like the above (online coaching, medical education). They may also, as a secondary outcome, support the future efforts of policy makers towards other conditions beyond cancer, as well as for disease prevention.

7.6 Data available and limitations

The data input provided, in the case of CRA, will be reported by breast cancer patients directly into a web platform, and will include information like the following:



Beyond this primary patient-reported input, there will be inferred input as well. More specifically, the guidance that patients receive will try to coach them into specific food behaviours. The adherence to this coaching, as well as the extent to which patients continue to provide input (a secondary measure of adherence) will be inferred from the ongoing patient data input as well. Therefore, these datasets will be the inputs for the causal analysis and forecasting module.

8. Conclusion and further work

In this document we have described the design and the specification of the causal analysis and forecasting framework in CrowdHEALTH platform. Given the basic principles of causality and forecasting this description will be used later on as the basis for the development of the causal analysis and forecasting prototype. The purpose of this toolkit is to explore causality and forecasting features important for policy makers and present the important health or life style factors that affect the population such as age, sex, race, previous disease diagnosis, drug prescription, geographical location and occupation. The main aim is to understand and present causality relationships that affect or have affected a disease development in relation to the public health policies for this disease.

Later in the project we are going to analyse further this framework and describe more analytically the causality and forecasting features based on:

1. The available data from the use case partners.
2. The description of the conditions and the causes of a specific disease development.(monitoring)

9. References

[1] Angrist JD, Imbens GW, Rubin DB (1996). Identification of causal effects using instrumental variables. Journal of the American Statistical Association 91:444—455.

[2] Dawid AP (1979). Conditional independence in statistical theory (with discussion). Journal of the Royal Statistical Society B 41:1—31.

[3] Dawid, A. P. (2003). Causal inference using influence diagrams: The problem of partial compliance (with discussion). In: Green PJ, Hjort NL, Richardson S, eds. Highly Structured Stochastic Systems. New York, NY: Oxford University Press, pp. 45—65.

[4] Flanders WD (2006). On the relation of sufficient component cause models with potential (counterfactual) models. European Journal of Epidemiology 21: 847—853.

[5] Greenland S, Brumback B (2002). An overview of relations among causal modelling methods. International Journal of Epidemiology 31:1030— 1037. Greenland S

[6] Greenland S, Lash TL, Rothman KJ (2008). Concepts of interaction. In: Rothman KJ, Greenland S, Lash TL, eds. Modern Epidemiology, 3rd edition. Philadelphia, PA: Lippincott Williams & Wilkins, pp. 71—83.

[7] Hernán MA, Hernández-Díaz S, Werler MM, Mitchell AA (2002). Causal knowledge as a prerequisite for confounding evaluation: an application to birth defects epidemiology. American Journal of Epidemiology 155:176— 184.

[8] Hernán MA, Robins JM (2006b). Instruments for causal inference: An epidemiologist's dream? Epidemiology 2006; 17(4): 360—372.

[9] Lash TL, Fox MP, Fink AK (2009). Applying Quantitative Bias Analysis to Epidemiologic Data. New York: Springer.

[10] Miettinen OS (1982). Causal and preventive interdependence: elementary principles. Scandinavian Journal of Work, Environment & Health 8:159— 168.

[11] Pearl J (2009). Causality: Models, Reasoning, and Inference, 2nd ed. New York: Cambridge University Press.

[12] Picciotto S, Hernán MA, Page J, Young JG, Robins JM (2012). Structural nested cumulative failure time models for estimating the effects of interventions. Journal of the American Statistical Association 107(499):886— 900.

[13] Richardson TS, Evans RJ, Robins JM (2010). Transparent parametrizations of models for potential outcomes. In: Bernardo JM, Bayarri MJ, Berger JO, Dawid AP, eds. Bayesian Statistics 9. Oxford University Press.

[14] M. A. Hernán, J. M. Robins. Causal Inference, Chapter 2, pages 11-29, 2017.

[15] J. M. Mooij, J. Peters, D. Janzing, J. Zscheischler and B. Scholköpf. Distinguishing cause from effect using observational data: methods and benchmarks. Journal of Machine Learning Research, pages 8-25, 2015.