FINDING CORRELATION BETWEEN CHRONICAL DISEASES AND FOOD CONSUMPTION FROM 30 YEARS OF SWISS HEALTH DATA LINKED WITH SWISS CONSUMPTION DATA USING FP-GROWTH FOR ASSOCIATION ANALYSIS

Jonas Baschung and Farshideh Einsele

Section of Business Information,
Bern University of Applied Sciences, Switzerland

ABSTRACT

Objective: The objective of the study was to link Swiss food consumption data with demographic data and 30 years of Swiss health data and apply data mining to discover critical food consumption patterns linked with 4 selected chronical diseases like alcohol abuse, blood pressure, cholesterol, and diabetes.

Design: Food consumption databases from a Swiss national survey menu CH were gathered along with data of large surveys of demographics and health data collected over 30 years from Swiss population conducted by Swiss Federal Office of Public Health (FOPH). These databases were integrated and Frequent Pattern Growth (FP-Growth) for the association rule mining was applied to the integrated database.

Results: This study applied data mining algorithm FP-Growth for association rule analysis. 36 association rules for the 4 investigated chronic diseases were found.

Conclusions: FP-Growth was successfully applied to gain promising rules showing food consumption patterns lined with lifestyle diseases and people's demographics such as gender, age group and Body Mass Index (BMI). The rules show that men over 50 years consume more alcohol than women and are more at risk of high blood pressure consequently. Cholesterol and type 2 diabetes is found frequently in people older than 50 years with an unhealthy lifestyle like no exercise, no consumption of vegetables and hot meals and eating irregularly daily. The intake of supplementary food seems not to affect these 4 investigated chronic diseases

KEYWORDS

Data Mining, Association Analysis, Apriori Algorithm, Diet & Chronical Diseases, Health Informatics.

1. INTRODUCTION

Chronical diseases increase in frequency across the globe, becoming an important public health problem even in developing countries. These diseases include obesity, hypertension (blood

David C. Wyld et al. (Eds): AIFZ, SIGPRO, NWCOM, CLSB, MLNLP, DTMN, BDIOT, CSITY - 2022
pp. 101-122, 2022. CS & IT - CSCP 2022 DOI: 10.5121/csit.2022.121609

pressure), heart disease, type 2 diabetes, cancer, mental disorders, and many others. They differ from the infectious diseases originated from malnutrition, also called communicable diseases (CD) due to their contagious, dispersive nature. Lifestyle diseases are therefore among the so-called NCD (non-communicable diseases) diseases. According to World Health Organization (WHO), the growing epidemic of chronic diseases afflicting both developed and developing countries are related to dietary and lifestyle changes [1].

Various researchers studied the relationship between nutritional habits and chronic diseases. Schulze et al in [2] discuss current knowledge on the associations between dietary patterns and multiple chronic diseases like cancer, heart disease, stroke, and type 2 diabetes. Their findings confirm that food-based prevention of chronic disease risk should prioritize fruits, vegetables, whole grains, fish and lower consumption of red and processed meats and sugar sweated drinks. Bocedi et al. state in [3] that an unhealthy lifestyle, like unbalanced diet, insufficient sleep, physical inactivity, smoking, alcohol abuse contributes the cause metabolic altercations which can lead to onset of NCDs. Some researchers in [4], [5], [6], [7], [8] studied particularly the impact of the Mediterranean diet, characterized by a high consumption of fruit, vegetables, extra virgin olive oil, cereals, legumes, and fish; a moderate intake of dairy products, eggs, and red wine; and a low intake of animal fats and red meat, as a correct approach to prevent NCDs. Di Marco et al. report in [9] specifically about pasta in Mediterranean diet and its antioxidant compounds like natural bioactive compounds play positive role in the protection of kidney cells from oxidative stress. Koch in [10] demonstrates that an optimal daily intake of antioxidants such as polyphenols and vitamins can counteract the onset of NCDs and to slow their progression. Chen et al. report in [11] that vitamin C (ascorbic acid) and E (tocopherols), are natural compounds that play a pivotal role in preventing the NCDs, mainly for their antioxidant activity. Vitamin C is a water-soluble vitamin, able to protect from the cellular damage exerted by harmful oxidative compounds. Noce et al. report in [12] how ω-3 polyunsaturated fatty acids play a cardioprotective role in male obesity secondary hypogonadism (MOSH) patients. Owen et al. in [13] report of their evaluation of the relationship between dietary quality scores and cardiometabolic risk in a group of older Australian adults, that a high intake of vegetables, grains, and non-processed red meat was associated with a better cardiometabolic risk profile.

Data Mining for chronic diseases prediction and prevention linked with nutritional habits have been explored by different researchers. Lee et al conducted a study using stepwise logistic regression (SLR) analysis, decision tree, random forest, and support vector machine as an alternative and complement to the traditional statistical approaches to identify the factors that affect the health-related quality of life (HRQoL) of the elderly with chronic diseases and to subse-quently develop from such factors a prediction model [14]. D. Qudsi and al. report in [15] from a study that aims to identify the potential benefits that data mining can bring to the health sector, using Indonesian Health Insurance company data as case study. Decision tree as a classification data mining method, was used to generate the prediction model by visualizing the tree to perform predictive analysis of chronic diseases. Z. Lei et al report in [16] of studying the relationship between nutritional ingredients and diseases such as diabetes, hypertension, and heart disease by using data mining methods. They have identified the first two or three nutritional ingredients in food that can benefit the rehabilitation of those diseases. R. McCabe et al. report in [17] of creating a simulation test environment using characteristic models of physician decision strategies and simulated populations of patients with type 2 diabetes, they state of employing a specific data mining technology that predicts encounter-specific errors of omission in representative databases of simulated physician patient encounters and test the predictive technology in an administrative database of real physician-patient encounter data. D.W. Haslam and W.P. James report in [18] of an investigation in a population - based sample of 1140 children performed to derive dietary patterns related to children's obesity status. Their findings reveal that Rules derived through a data mining approach revealed the detrimental influence of the increased

consumption of fried food, delicatessen meat, sweets, junk food and soft drinks. K. Lange et al. state in [19] that big data studies may ultimately lead to personalized genotype-based nutrition which could permit the prevention of diet-related diseases and improve medical therapy. A. Hearty and M. Gibney evaluate the usability of supervised data mining methods as ANNs and decision trees to predict an aspect of dietary quality an aspect of dietary quality based on dietary intake with a food-based coding system and a novel meal-based coding system [20]. A. von Reusten et al. used data from 23 531 participants of the EPIC-Potsdam study to analyze the associations between 45 single food groups and risk of major chronic diseases, namely, cardiovascular diseases (CVD), type 2 diabetes and cancer using multivariable-adjusted Cox regression. Their results show that higher intakes of low-fat dairy, butter, red meat, and sauce were associated with higher risks of chronic diseases [21]. E. Yu et al. demonstrate in [22] the usability of supervised data mining methods to extract the food groups related to bladder cancer. Their results show that beverages (non-milk); grains and grain products; vegetables and vegetable products; fats, oils, and their products; meats and meat products were associated with bladder cancer risk.

As a proof of concept, we conducted a preliminary study [23], in which we used a big database gained from a grocery store chain over a certain period along with associated health data of the same region. Association rule mining was successfully used to describe and predict rules linking food consumption patterns with lifestyle diseases. Additionally, we conducted a further study using a medium-sized real-world health and nutritional data from Swiss population and gained interesting rules which showed the link between nutritional habits and chronical diseases. [24] and later another study, in which we used the same national Swiss dietary survey with a five times larger dataset (collected over 25 years) from the national Swiss health survey including demographical information [25]. Based on the finding of the previous studies, where it used the pure Apriori algorithm which resulted that some critical health-related dietary features were pruned out early in course of data mining, we have applied the Weighted Association Mining Rules (WARM) analysis to the latter study.

In This study, we have enlarged our health date from 1991 to 2017 and used Frequent Pattern Growth (FP-Growth) algorithm to gain rules that show the link between Swiss nutritional habits and chronical diseases. Additionally, in this study we added BMI, age group and gender to our candidate patterns, which helped us to gain more specific association rules which contain demographical information when assessing the relationship between chronic diseases and nutrition.

2. SELECTION, CLEANING, TRANSFORMATION OF THE DATABASES

The following formatting rules must be followed strictly.

2.1. Selection

The data comes from the national surveys menuCH and the health survey that were carried out in Switzerland. The national food survey menuCH (BLV, Federal Office for Food Safety and Veterinary 2020) was carried out for the first time from January 2014 to February 2015. Over 2000 people living in Switzerland were asked about their eating habits and food consumption. The data resulting from the survey is the first representative, national nutritional survey data available in Switzerland from BLV. The second data source comprises health data on the state of health and health-related behavior of the Swiss resident population over a period of 30 years. The Federal Statistical Office (2021) has been collecting health data from the population living in Switzerland every five years using a writ-ten and telephone questionnaire. As part of this study,

representative data from around 85,000 people from 1992, 1997, 2002, 2007, 2012 are available. This data has already been pre-cleaned, attributes have been partially selected from the database and the data has been already transformed as reported in [25]. In addition to this, the author has added the health data of 2017 to the health database.

2.2. Cleaning

2.2.1. Cleaning menuCH database

MenuCH database was remained untouched, as described in the study of Mewes et al. [24].

2.2.2. Cleaning health database

From the health data, all records that contained blank or missing survey responses were removed so that only complete data sets are used for analysis. Data cleaning resulted in a significant reduction in the number of usable records for all diseases. For disease alcohol consumption, for example, the number of records was reduced from the original 108,267 to 12,685 records.

For this purpose, all health responses of all persons (always 8 responses per person) including the demographic characteristics were exported from the SQL database for each disease examined, loaded into an Excel, and examined for completeness. The incomplete data sets were eliminated and a new file with complete data sets was then exported from Excel to CSV and used for the investigation in Python. Finally, we have built categories for the 4 selected chronical diseases as follows:

Categories Alcohol:

- 0-17 gr. Alcohol consumption daily,
- 18-22 gr. alcohol consumption daily
- 23-28 gr. Alcohol consumption daily
- more than 28 gr. Alcohol consumption daily

Categories Blood pressure

- not medically assessed normal
- medically judged normal
- not medically judged too low
- medically judged too high
- not medically judged too high
- medically judged too low

Categories Cholesterol

- not medically judged normal
- medically judged normal
- medically judged too high
- not medically assessed too high

Categories Diabetes

- not medically assessed, no diabetes
- medically assessed no diabetes

- medically assessed diabetes
- not medically assessed, diabetes

2.3. Transformation

For the integration of the nutrition and health databases, a third person profile table had to be created, which connects the person profile tables of the nutrition database and the health database. Six attributes were selected which were available in both databases for the personal description:

- Gender (m / f)
- age group (15-29 / 30-39 / 40-49 / 50-64 / 65+)
- Household size (1/2/3/4/5 / 6+)
- Marital status (single / married or registered / widowed / divorced / other)
- Language (de / fr / it).

The selected attributes and their categories resulted in 720 different categories. The PersonIDs in the Menu-CH database and the PersonIDs in the Health database were each assigned to a person category in the PersonProfil table as shown in Fig. 1.

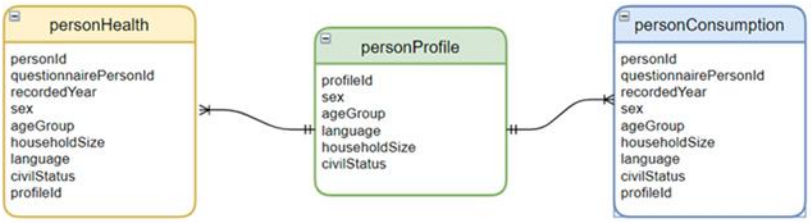

Figure 1. Linking table

Moreover, to work with the 4 selected chronical diseases more efficiently, we decided to add a dimensional scheme to our previously relational database reported in (Lustenberger, 2021) and build 5 dimensions around the fact table personhealth. These dimensions are the 4 selected chronical diseases Alcohol, Blood Pressure, Cholesterol, and diabetes type 2 along with BMI dimension. (see Fig. 2)

2.4. Integration

Fig. 2 shows the scheme of the integrated database with the personProfile as the central link, the structure of the menu-CH data and the connection to health database containing health data from 1991 to 2017, which is a hybrid relational & dimensional scheme containing the relational tables (bottom left) and the dimensional tables (top right). Fig. 2 shows our new hybrid database scheme.

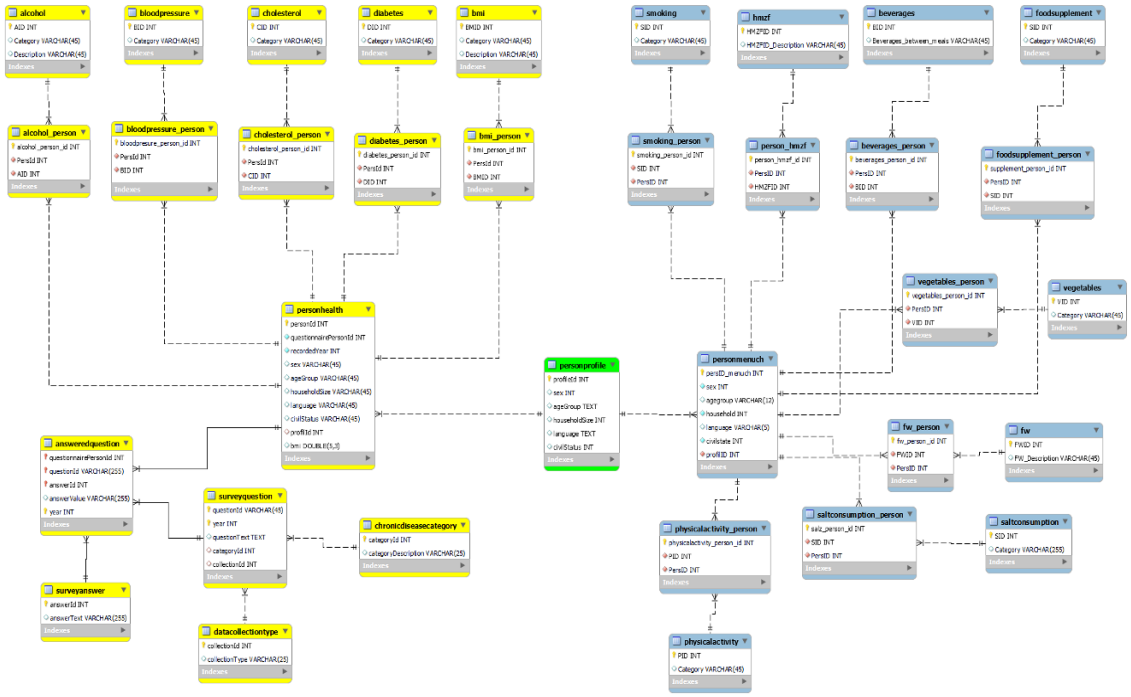

Figure 2. Scheme of the integrated hybrid (relation & dimensional) database

3. ASSOCIATION ANALYSIS WITH FP-GROWTH ALGORITHM

Since our database has grown compared to the previous studies, we decided to use FP-Growth algorithm in this study. FP Growth has been proven to be more performant than Apriori algorithm as a frequent pattern is generated without the need for candidate generation in contrast to Apriori algorithm. FP growth algorithm represents the database in the form of a tree called a frequent pattern tree or FP tree. This tree structure will maintain the association between the itemsets. The database is fragmented using one frequent item. This fragmented part is called “pattern fragment”. The itemsets of these fragmented patterns are analyzed. Thus, with this method, the search for frequent itemsets is reduced accordingly. We have created a total of four Python scripts, in which the analyzes for the four chronic diseases alcohol consumption, high blood pressure, cholesterol and diabetes were carried out. Common functions, which are executed by all four Python scripts, were refactored into a separate "Helper" class. In order to be able to process the data to be analyzed using the methods provided by «mlxtend.frequent_patterns», these should have been available in a one-dimensional, encoded numpy array. The Excel file resulting from the data cleansing was converted into an importable CSV file. The CSV file was loaded directly into the respective Python script and then converted.

```python
# Load file that contains all categorized data with demographic categorized
bloodpressure = pd.read_csv("data/bloodpressure_cleaned_demographic.csv", encoding="utf-8", sep=';',
    dtype={"bloodpressure_person_id": int, "bloodpessure_demographic": "string", "PhysicalActivity": "string",
    "Haupmahlzeiten": "string", "Gemüseverzehr": "string", "Smoking": "string", "Nahrungsergänzungsmittel": "string",
    "Salzkonsum": "string", "Anz warm Mahlzeiten": "string", "Beverages_between_meals": "string"})

# Delete column persID because not need
del bloodpressure["bloodpressure_person_id"]

#Convert demograpgic data to string
bloodpressure['bloodpessure_demographic'] = bloodpressure['bloodpessure_demographic'].astype(str)

#Convert data to numpy array
bloodpressure = Helpers.importedDFToListToNumpy(bloodpressure)

# Function to improve the imported df to work with mlextend
def importedDFToListToNumpy(df):
    list = df.values.tolist()
    prettynArray = np.array(list)
    return prettynArray

```

Figure 3. Python script for preparation of data

Converting it to a one-dimensional numpy array required several steps. The characteristics sometimes contain longer answers separated by commas, such as "Water, coffee/tea, SFGEI, Light, Milchg." when consuming drinks. This caused Numpy to convert the imported dataframe into a multidimensional array. For this reason, a function was created which successfully converts the dataframe into a numpy array. In a next step, the existing array is converted from a "pandas" dataframe into a boolean numpy array, which can be used directly to search for frequent item sets.

```
array([[ True, False,  True,  True, False,  True],
       [ True, False,  True, False, False,  True],
       [ True, False,  True, False, False, False],
       [ True,  True, False, False, False, False],
       [False, False,  True,  True,  True,  True],
       [False, False,  True, False,  True,  True],
       [False, False,  True, False,  True, False],
       [ True,  True, False, False, False, False]], dtype=bool)
```

Figure 4. Numpy array

The occurrence of each feature in each transaction is stored in this array, which makes it easier to find frequent item sets. The Boolean data frame that is now available can be passed to the FP-Growth function (Fig. 6) to generate frequent item sets.

```python
#Run fpGrwoth for whole dataset
resAlcohol = Helpers.runFPGrowth(dfAlcohol, 0.001)

# Running fpgrwoth the with boolean encoded numpy array and minsupport
def runFPGrowth(earray, min_support):
    print("FPGROWTH is processing, with minSupport: " + str(min_support)
    res = fpgrowth(earray, min_support=min_support, use_colnames=True)
    return res

```

Figure 5. Python script to run the data frame

The execution of the FP growth algorithm was carried out in several iterative runs for each chronic disease, in which the results are examined with different support values. It was started with a support of 1%. Since the number of sick people is small in relation to the healthy people in the existing data set, disease characteristics appear less often in the results. Association rules

could be set up from the frequently generated item sets. For this purpose, the Python code was implemented and tested. Afterwards, association rules were generated iteratively for each of the four chronic diseases alcohol consumption, hypertension, cholesterol, and diabetes. The rules were first created based on the parameters support, confidence, and lift, then filtered, sorted and exported. Once exported, they were further examined, sorted, and filtered in Excel.

```
# creating association rules based on lift, confidence or support
resAssociationAlcohol = Helpers.createAssociationRules(resAlcohol, "confidence", 0.01)
#resAssociationAlcohol = Helpers.createAssociationRules(resAlcohol, "lift", 0.1)
#resAssociationAlcohol = Helpers.createAssociationRules(resAlcohol, "support", 0.01)

#Create association rules
def createAssociationRules(resultData, metric, min_threshold):
    print("Creating Association Rules with metric: " + str(metric) + " with threshold of: " + str
    rules = association_rules(resultData, metric=metric, min_threshold=min_threshold)
    return rules
```

Figure 6. Python script to create association rules using FP-Growth algorithm

Fig. 6 shows how wanted and unwanted conclusions are stored in the variables search for Consequents and not needed Consequents (neither of which are shown completely). These were filtered out of the data frame with the generated association rules using regular expressions. After filtering, the association rules were each exported in a single CSV file for the parameters support, confidence, and lift. The CSV files could then be imported into Excel files and the data could be further analyzed there.

4. GAINED ASSOCIATION RULES

A total of 36 new association rules, 10 rules for alcohol abuse, 9 rules for blood pressure (hypertension), 9 rules for cholesterol and 8 rules for diabetes containing the relevant diet and the newly added demographics such as gender, age group and BMI of Swiss residents could be found. We have grouped them based on their highest lift, support, and confidence. The highest lift shows xxx, the highest support shows xxx and the highest confidence shows xxx.

4.1. Association rules for alcohol abuse

4.1.1. Association rules with the highest lift

Rule 1

1.2% of the sample consumed more than 28 gr of alcohol daily, were men, older than 65 years and had a BMI between 18.5-24.9 and had the following characteristics:

- They rarely move (0 days a week)
- They use salt without additives, without re-salting at home
- They regularly eat breakfast, lunch, and dinner

This rule meets in 4.1% of the sample (confidence) and has a lift of 3.5

Rule 2

1.2% of the sample consumed more than 28 gr of alcohol daily, were men, older than 65 years and had a BMI between 18.5-24.9 and had the following characteristics:

- They rarely move (0 days a week)
- They use salt without additives, without re-salting at home
- They regularly eat breakfast, lunch, and dinner
- They never to rarely eat a hot meal

This rule is true in 2% of the sample (confidence) and has a lift of 1.6

Rule 3

1.5% of the sample consume daily 0-17 gr. Alcohol, are women, between 15-29 years old with a BMI below 18.5 and have the following characteristics:

- They regularly exercise between five to seven times a week
- They never to rarely eat a hot meal
- They never to rarely (0-3x per month) eat vegetables
- They use salt without additives, without re-salting at home
- They drink water, coffee, or tea between meals
- They regularly eat breakfast, lunch, and dinner

This rule is true in 70% of the sample (confidence) and has a lift of 63

Rule 4

1.5% of the sample consume daily 0-17 gr. Alcohol, are women, between 50-64 years old with a BMI below 18.5 and have the following characteristics:

- They regularly exercise between five to seven times a week
- They never to rarely eat a hot meal
- They never to rarely (0-3x per month) eat vegetables
- They use salt without additives, without re-salting at home
- They drink water, coffee, or tea between meals
- They regularly eat breakfast, lunch, and dinner

This rule is true in 59% of the sample (confidence) and has a lift of 44

Rule 5

1.3% of the sample consume daily 0-17 gr. Alcohol, are men, between 50-64 years old with a BMI below 18.5 and have the following characteristics:

- They rarely move (0 days a week)
- They never to rarely eat a hot meal
- They never to rarely (0-3x per month) eat vegetables
- They use salt without additives, without re-salting at home
- They drink SFGEI (assumption: sugar-free beverages) between meals.
- They regularly eat breakfast, lunch, and dinner

This rule is true in 51% of the sample (confidence) and has a lift of 36

4.1.2. Association rules with the highest support

Rule 6

7.3% of the sample consume daily 0-17 gr. Alcohol, are women, between 50-64 years old with a BMI 18.5-24.9 and have the following characteristics:

- They never to rarely eat a hot meal
- They never to rarely (0-3x per month) eat vegetables
- They use salt without additives, without re-salting at home
- They regularly eat breakfast, lunch and dinner
- They move irregularly (1-4 days)

This rule is true in 14% of the sample (confidence) and has a lift of 1.92

Rule 7

7.3% of the sample consume daily 0-17 gr. Alcohol, are women, between 50-64 years old with a BMI 18.5-24.9 and have the following characteristics:

- They never to rarely eat a hot meal
- They never to rarely (0-3x per month) eat vegetables
- They do not take dietary supplements
- They regularly eat breakfast, lunch, and dinner

This rule is true in 9.4% of the sample (confidence) and has a lift of 1.28

4.1.3. Association rules with the highest confidence

Rule 8

4.89% of the sample consume daily 0-17 gr. Alcohol, are women, 65 years and older with BMI 25-29.9 and have the following characteristics:

- They never to rarely eat a hot meal
- They never eat vegetables irregularly (1-2x week)
- They do not take dietary supplements
- They use salt without additives, without re-salting at home
- They move irregularly (1-4 days per week)
- They regularly eat breakfast, lunch, and dinner
- They drink coffee/tea and milk drinks (between meals)

This rule is true in 100% of the sample (confidence) and has a lift of 20

Rule 9

7.34% of the sample consume daily 0-17 gr. Alcohol, are women, between 50-64 years old with a BMI 18.5-24.9 and have the following characteristics:

- They never to rarely eat a hot meal
- They never to rarely eat vegetables
- They regularly eat breakfast, lunch, and dinner

- They drink coffee/tea and SFGEI (between meals).
- They do not take dietary supplements
- They use salt without additives, without re-salting at home
- They move irregularly (1-4 days per week)

This rule is true in 61% of the sample (confidence) and has a lift of 8.4

Rule 10

Rules with high alcohol consumption have a lower confidence because they occur less often, nevertheless two rules with the highest confidence from this category were established:

1.19% of the sample consumed more than 28 gr of alcohol daily, were men, 65 years and older, and had a BMI between 18.5-24.9 and had the following characteristics:

- They never to rarely eat a hot meal
- They never to rarely eat vegetables
- They regularly eat breakfast, lunch, and dinner
- They use salt without additives, without re-salting at home

This rule meets in 1.95% of the sample (confidence) and has a lift of 1.6.

4.2. Association rules for blood pressure (hypertension)

In our health database, we had records with medically assessed and medically not assessed hypertension. Table 1 shows this distribution in the database. However, we applied data mining on records which were medically assessed.

Table 1. Distribution of medically assessed records for blood pressure

Expression	Num. of transactions	Share in %
not medically assessed normal	28'889	65,69
medically assessed normal	6'662	15,15
not medically assessed too low	4'942	0,35
medically judged too high	1'812	4,12
not medically assessed too high	1'520	3,46
medically judged too low	154	11,24
Total	43'979	100

4.2.1. Association rules with highest lift

Rule 1

1.27% of the sample have medically judged normal blood pressure, are women, 65 years and older, have a BMI of 25-29.9, and have the following characteristics:

- They move irregularly (1-2x / week)
- They use salt without additives, without re-salting at home
- They do not take dietary supplements
- They drink coffee/tea, milk drinks (between meals).

- They regularly eat breakfast, lunch, and dinner
- They never to rarely eat hot meals
- They do not smoke and did not smoke before
- They prepare irregularly (1-2x / week)) vegetables

This rule is true in 32% of the sample (confidence) with a lift of 17.8

Rule 2

0.23% of the sample have medically assessed hypertension, are women, 65 years and older, have a BMI of 18.5 - 24.9, and have the following characteristics:

- They do not take dietary supplements
- They use salt without additives, without re-salting at home
- They never to rarely eat vegetables irregularly (0-3x / month)
- They drink coffee/tea (between meals)
- They eat hot meals irregularly (4-7x / week)
- They exercise regularly (5-7 days per week)
- They do not smoke and did not smoke before
- They regularly eat breakfast, lunch, and dinner

This rule is true (confidence) in 9% of the sample with a lift of 22.3

Rule 3

0.13% of the sample have medically assessed hypertension, are men, 65 years and older, have a BMI of >= 30, and have the following characteristics:

- They do not take dietary supplements
- They use salt without additives, without re-salting at home
- They never to rarely eat vegetables irregularly (0-3x / month)
- They drink coffee/tea
- They regularly eat breakfast, lunch, and dinner
- They never to rarely eat hot meals (0-3x / week)

This rule is true in 2.6% of the sample (confidence) with a lift of 19.

4.2.2. Association rules with highest support

Rule 4

0.7% of the sample have medically assessed hypertension, are men, 65 years and older, have a BMI of 18.5-24.9, and have the following characteristics:

- They do not take dietary supplements
- They use salt without additives, without re-salting at home
- They never to rarely eat vegetables (0-3x / month)
- They drink coffee/tea (between meals)
- They regularly eat breakfast, lunch, and dinner
- They never to rarely eat hot meals (0-3x / week)

- They do not smoke and have not smoked before

This rule is true in 1.2% of the sample (confidence) with a lift of 1.7

Rule 5

0.33% of the sample have medically assessed hypertension, are women, 65 years and older, have a BMI of 25-29.9, and have the following characteristics:

- They do not take dietary supplements
- They use salt without additives, without re-salting at home
- They eat vegetables irregularly (1-2x / week)
- They move irregularly (1-4 days / week)
- They drink coffee/tea and milk drinks (between meals)
- They regularly eat breakfast, lunch, and dinner
- They never to rarely eat hot meals (0-3x / week)
- They do not smoke and did not smoke before

This rule is true in 8.4% of the sample (confidence) with a lift of 17

Rule 6

2.74% of the sample have medically assessed normal blood pressure, are men, 65 years and older, have a BMI of 24.9-29.9, and have the following characteristics:

- They do not take dietary supplements
- They regularly eat breakfast, lunch, and dinner
- They never to rarely eat hot meals
- They never - rarely eat vegetables (0-3x month)

This rule is true in 4.5% of the sample (confidence) with a lift of 1.5

4.2.3. Association rules with highest confidence

Rule 7

2.9% of the sample have medically assessed normal blood pressure, are men, 65 years and older, have a BMI of 24.9-29.9, and have the following characteristics:

- They drink alcoholic beverages
- They rarely to never move
- They rarely - never eat vegetables
- They do not take dietary supplements

This rule is true in 36% of the sample (confidence) with a lift of 12

Rule 8

0.41% of the sample have medically assessed high blood pressure, are women, 65 years and older, have a BMI of 18.5-24.9, and have the following characteristics:

- They eat vegetables irregularly (1-2x / week)
- They eat hot meals irregularly (4-7x/week)
- They move regularly (5-7 days / week)
- They use salt without additives, without re-salting at home

This rule is true in 9.2% of the sample (confidence) with a lift of 22

Rule 9

0.41% of the sample have medically assessed high blood pressure, are women, 65 years and older, have a BMI of 18.5-24.9, and have the following characteristics:

- They do not take dietary supplements
- They use salt without additives, without re-salting at home
- They eat vegetables irregularly (1-2x / week)
- They move irregularly (1-2x / week)
- They drink coffee/tea and milk drinks (between meals)
- They never to rarely eat hot meals (0-3/month)
- They do not smoke and did not smoke before

This rule is true in 8.4% of the sample (confidence) with a lift of 17

4.3. Association rules for cholesterol

In our health database, we had records with medically assessed and medically not assessed cholesterol. Table 2 shows this distribution in the database. However, we applied data mining on records which were medically assessed.

Table 2. Distribution of medically assessed records for cholesterol

Expression	No. of transactions	Share in %
not medically assessed normal	26'812	80,57
medically assessed normal	3'936	11,83
medically judged too high	1'364	4,10
not medically assessed too high	1'164	3,50
Total	33'276	100

4.3.1. Association rules with highest lift

Rule 1

0.28% of the sample had medically assessed normal cholesterol, were men, 65 years and older, had a BMI of 25-29.9, and had the following characteristics:

- They drink water and alcoholic beverages
- They do not take dietary supplements
- They regularly eat breakfast, lunch, and dinner
- They do not smoke and have not smoked before
- They rarely to never eat vegetables (0-3x/month)
- They rarely to never eat hot meals (0-3x / week)
- They use salt without additives, and regular re-salting at home (1-5/10 meals).

This rule is true in 0.15% of the sample (confidence) with a lift of 108

Rule 2

0.27% of the sample had medically assessed high cholesterol, were women, 65 years and older, had a BMI of 18.5-24.9, and had the following characteristics:

- They use salt without additives, without regularly re-salting at home
- They drink water, tea/coffee (between meals)
- They do not take dietary supplements
- They regularly eat breakfast, lunch, and dinner
- They eat hot meals irregularly (4-7x / week)
- They eat vegetables irregularly (1-2x/week)

This rule is true in 4.7% of the sample (confidence) with a lift of 17

Rule 3

0.13% of the sample had medically assessed high cholesterol, were women, 65 years and older, had a BMI of >=30, and had the following characteristics:

- They use salt without additives, without regularly re-salting at home
- They drink water, tea/coffee (between meals)
- They do not take dietary supplements
- They regularly eat breakfast, lunch, and dinner
- They rarely to never eat hot meals (0-3x / week)
- They eat vegetables regularly (>2 /week)
- They do not smoke and did not smoke before

This rule is true in 11% of the sample (confidence) with a lift of 79

Rule 4

0.13% of the sample had medically assessed high cholesterol, were men, 65 years and older, had a BMI of >=30, and had the following characteristics:

- They never to rarely eat vegetables (0-3 / month)
- They use salt without additives, without regularly re-salting at home
- They rarely to never eat hot meals (0-3x / week)
- They drink coffee/tea (between meals)

This rule is true in 2.4% of the sample (confidence) with a lift of 17

4.3.2. Association rules with highest support

Rule 5

2.4% of the sample had medically assessed normal cholesterol, were men, 65 years and older, had a BMI of 18.5-24.9, and had the following characteristics:

- They use salt without additives, without regularly re-salting at home
- They do not take dietary supplements

- They regularly eat breakfast, lunch, and dinner
- They do not smoke and did not smoke before
- They rarely to never eat vegetables (0-3x/month)
- They rarely to never eat hot meals (0-3x / week)

This rule is true in 4.1% of the sample (confidence) with a lift of 1.7

4.3.3. Association rules with highest confidence

Rule 6

2.4% of the sample had medically assessed normal cholesterol, were women, 65 years and older, had a BMI of 25-29.9, and had the following characteristics:

- They use salt without additives, without regularly re-salting at home
- They move regularly (5-7 days / week)
- They do not take dietary supplements
- They regularly eat breakfast, lunch, and dinner
- They do not smoke and did not smoke before
- They rarely to never eat hot meals (0-3x / week)
- They eat vegetables irregularly (1-2x/week)
- They drink tea/coffee and milk drinks (between meals)

This rule is true in 20.9% of the sample (confidence) with a lift of 15

Rule 7

0.11% of the sample had medically assessed high cholesterol, were men, 65 years and older, had a BMI of 18.5-24.9, and had the following characteristics:

- They use salt without additives, without regularly re-salting at home
- They move regularly (5-7 days / week)
- They never to rarely eat vegetables (0-3x/month)
- They rarely to never eat hot meals (0-3x / week)
- They drink alcoholic beverages (between meals)

This rule is true in 6.06% of the sample (confidence) with a lift of 11

Rule 8

0.27% of the sample had medically assessed high cholesterol, were women, 65 years and older, had a BMI of 18.5-24.9, and had the following characteristics:

- They use salt without additives, without regularly re-salting at home
- They drink water, coffee/tea (between meals)
- They do not take dietary supplements
- They regularly eat breakfast, lunch, and dinner
- They eat hot meals irregularly (4-7x / week)
- They do not smoke and did not smoke before
- They eat vegetables irregularly (1-2x / week)

This rule is true in 4.7% of the sample (confidence) with a lift of 17

4.4. Association rules for diabetes

In our health database, we had records with medically assessed and medically not assessed diabetes. Table 3 shows this distribution in the database. However, we applied data mining on records which were medically assessed.

Table 3. Distribution of medically assessed records for diabetes

Expression	No. of transactions	Share in %
not medically assessed, no diabetes	26'648	94,08
medically assessed no diabetes	1174	4,14
medically assessed diabetes	328	1,16
not med. assessed, diabetes	174	0,61
Total	28'324	100

4.4.1. Association rules with highest lift

Rule 1

0.21% of the sample have medically assessed diabetes, are men, 65 years and older, have a BMI of 18.5-24.9, and have the following characteristics:

- They never to rarely eat vegetables (0-3x / month)
- They never to rarely eat hot meals (0-3x / week)
- They do not take dietary supplements
- They drink alcoholic beverages (between meals)
- They consume salt without addition never re-salt at home

This rule is true in 1.9% of the sample (confidence) and a lift of 8.9

Rule 2

0.15% of the sample have medically assessed diabetes, are men, 50 to 64 years old, have a BMI of 18.5-24.9, and have the following characteristics:

- They never to rarely eat vegetables (0-3x / month)
- They never to rarely eat hot meals (0-3x / week)
- They do not take dietary supplements
- They drink alcoholic beverages (between meals)
- They consume salt without addition never re-salt at home
- They do not smoke and did not smoke before

This rule is true in 0.21% of the sample (confidence) and a lift of 1.4

Rule 3

0.26% of the sample did not have diabetes as medically assessed, were women, 65 years and older, had a BMI of 18.5-24.9, and had the following characteristics:

- They do not take dietary supplements
- They eat vegetables irregularly (1-2x / week)

- They move regularly (5-7 days / week)
- They consume salt without addition never re-salt at home
- They eat hot meals irregularly (4-7x / week)

This rule is true in 5.7% of the sample (confidence) and a lift of 21

Rule 4

- 0.21% of the sample did not have diabetes as medically assessed, were men, 65 years and older, had a BMI of >= 30 and had the following characteristics:
- They never to rarely eat vegetables (0-3x / month)
- They move irregularly (1-4 days / week)
- They never to rarely eat hot meals (0-3x / week)
- They consume salt without addition never re-salt at home
- They consume coffee/tea (between meals)

This rule is true in 3.9% of the sample (confidence) and a lift of 18.7

4.4.2. Association rules with highest support

Rule 5

0.21% of the sample have medically assessed diabetes, are men, 65 years and older, have a BMI of 18.5-24.9, and have the following characteristics:

- They never to rarely eat vegetables (0-3x / month)
- They never to rarely eat hot meals (0-3x / week)
- They do not take dietary supplements
- They regularly eat breakfast, lunch, and dinner
- They do not smoke and did not smoke before

This rule is true in 0.29% of the sample (confidence) and a lift of 1.4

Rule 6

1% of the sample are medically assessed not to have diabetes, are men, 65 years and older, have a BMI of 18.5-24.9, and have the following characteristics:

- They never to rarely eat vegetables (0-3x / month)
- They never to rarely eat hot meals (0-3x / week)
- They do not take dietary supplements
- They regularly eat breakfast, lunch, and dinner
- They do not smoke and did not smoke before
- They consume salt without addition never re-salt at home

This rule is true in 1.6% of the sample (confidence) and a lift of 1.6

4.4.3. Association rules with the highest confidence

Rule 7

0.22% of the sample have medically assessed diabetes, are men, 65 years and older, have a BMI of 18.5-24.9, and have the following characteristics:

- They never to rarely eat vegetables (0-3x / month)
- They never to rarely eat hot meals (0-3x / week)
- They do not take dietary supplements
- They drink alcoholic beverages (between meals)

This rule is true in 1.9% of the sample (confidence) and a lift of 8.9

Rule 8

0.11% of the sample have medically assessed diabetes, are women, 65 years and older, have a BMI of 25-29.9, and have the following characteristics:

- They consume salt without addition never re-salt at home
- They do not take dietary supplements
- They regularly eat breakfast, lunch, and dinner
- They do not smoke and did not smoke before

This rule is true in 0.14% of the sample (confidence) and a lift of 1.2

Rule 9

0.51% of the sample did not have diabetes as medically assessed, were women, 65 years and older, had a BMI of 25.0-29.9, and had the following characteristics:

- They do not take dietary supplements
- They drink coffee/tea and milk drinks (between meals)
- They do not smoke and did not smoke before
- They never to rarely eat hot meals (0-3x / week)
- They consume salt without addition never re-salt at home
- They regularly eat breakfast, lunch, and dinner

This rule is true in 6.7% of the sample (confidence) and a lift of 12

5. CONCLUSION AND FUTURE WORK

In this study, we applied FP-Growth to find association rules that demonstrate the relationship between nutritional habits and four chronic diseases such as alcohol abuse, blood pressure, cholesterol, and diabetes along with the corresponding demographics. Our data base includes health data from multiple surveys over 30 years (1992-2017) from tens of thousands of Swiss population and nutrition data from the first Swiss nationwide nutritional survey (2014-2015). In the previous studies (Mewes, Einsele, 2020) and (Lustenberger, Einsele, 2021), we have dealt with smaller health databases and applied Apriori algorithm to extract association rules. The nutritional data in the previous study and the current study remains unchanged. Since the health database has grown significantly in this study, we chose FP-Growth and used a Python's machine learning library to be more performant and efficient comparing to Apriori algorithm in the

previous studies. Further improvement in the current study is that we have added demographic information about gender, age group and Body Mass Index (BMI) to the list of itemsets to gain more accurate association rules from our integrated database.

The study shows that concerning alcohol abuse men over 65 years that have a normal BMI, who overconsume alcohol daily, rarely move and rarely eat hot meals. Furthermore, middle aged men between 54 and 65 years with normal BMI, who do not overconsume alcohol, have a similar lifestyle to the previous group like making no exercise and rarely eating hot meals and vegetables. Women, on the contrary, who are younger than 64 years old with an ideal BMI don't overconsume alcohol. These women exercise regularly but rarely eat hot meal or vegetables but eat 3 times a day. Hence, although their lifestyle is similar to the men of same age, women tend to less over consume alcohol.

In addition to this, women older than 65 years with normal to high BMI with hypertension, do not smoke, exercise weekly, eat regularly but rarely eat hot meals or vegetables. Men older than 65 with a normal and high BMI with hypertension eat regularly but rarely hot meals or vegetables as well. Additionally, our rules show that intake of dietary supplements does not reduce blood pressure.

Furthermore, women older than 65 years with a normal BMI who have high cholesterol, eat hot meals and vegetables irregularly. Additionally, women over 65 years with a high BMI, eat vegetables both regularly and irregularly, don't smoke, eat rarely hot meals but 3 meals a day. Finally, Men over 65 years old with a normal BMI who have high cholesterol, rarely eat vegetables, and hot meals but drink daily alcoholic beverages. This is in accordance with our found rules about alcohol consume, which shows that older men tend to more overconsume alcohol, and this could result to high cholesterol as well.

Finally, our rules show that mostly men and women over 50 years show diabetes type 2. Which can be a result of years of unhealthy lifestyles. According to our found rules, men over 50 years with normal to high BMI, rarely eat vegetables and hot meals and consume alcoholic beverages daily. Interestingly additional rules show that the same age group no matter which gender who eat three meals a day but eat rarely hot meals and vegetables have no diabetes. This could be an important hint that eating three meals a day by people older than 50 could reduce the risk of diabetes 2. For future work, it is essential to expand the data base. The data base is unevenly distributed with two different data sources menuCH and SGB (health data). With 2'000 persons from the nutrition survey (menuCH) and a total of about 120'000 persons from the SGB. Firstly, we need to extract more nutrition and lifestyle data as possible from SGB. As of 2017, SGB also asked questions about tobacco use, additional eating, and physical activity behaviors. Momentarily, we are developing a shopping basket nutritional data base to increase the quantity of nutritional records as well as to mitigate the wishful thinking behavior, which is an important biased factor in the common surveys.

REFERENCES

[1] WHO, 2003. Diet, Nutrition, and the Prevention of Chronic Diseases. Report of a Joint WHO/FAO Ex-pert Consultation. *World Health Organization aper templates.*

[2] M. B. Schulze et al, Food based dietary patterns and chronic disease prevention, *BMJ 2018*; 361 doi: https://doi.org/10.1136/bmj.k2396, 13 June 2018

[3] Di Daniele, N.; Noce, A.; Vidiri, M.F.; Moriconi, E.; Marrone, G.; Annicchiarico-Petruzzelli, M.; D'Urso, G.; Tesauro, M.; Rovella, V.; De Lorenzo, A. Impact of Mediterranean diet on metabolic syndrome, cancer and longevity. *Oncotarget 2017*, 8, 8947–8979.

[4] De Lorenzo, A.; Noce, A.; Bigioni, M.; Calabrese, V.; Della Rocca, D.G.; Di Daniele, N.; Tozzo, C.; Di Renzo, L. The effects of Italian Mediterranean organic diet (IMOD) on health status. *Curr. Pharm. Des. 2010*, 16, 814–824.

[5] Andreoli, A.; Lauro, S.; Di Daniele, N.; Sorge, R.; Celi, M.; Volpe, S.L. Effect of a moderately hypoenergetic Mediterranean diet, and exercise program on body cell mass and cardiovascular risk factors in obese women. Eur. J. Clin. Nutr. 2008, 62, 892–897.

[6] Di Daniele, N.; Di Renzo, L.; Noce, A.; Iacopino, L.; Ferraro, P.M.; Rizzo, M.; Sarlo, F.; Domino, E.; De Lorenzo, A. Effects of Italian Mediterranean organic diet vs. low-protein diet in nephropathic patients according to MTHFR genotypes. *J. Nephrol. 2014*, 27, 529–536.

[7] Noce, A.; Marrone, G.; Urciuoli, S.; Di Daniele, F.; Di Lauro, M.; Pietroboni Zaitseva, A.; Di Daniele, N.; Romani, A. Usefulness of Extra Virgin Olive Oil Minor Polar Compounds in the Management of Chronic Kidney Disease Patients. *Nutrients 2021*, 13, 581.

[8] Noce, A.; Fabrini, R.; Bocedi, A.; Di Daniele, N. Erythrocyte glutathione transferase in uremic diabetic patients, *Acta Diabetol. 2015*, 52, 813–815.

[9] Di Marco, F.; Trevisani, F.; Vignolini, P.; Urciuoli, S.; Salonia, A.; Montorsi, F.; Romani, A.; Vago, R.; Bettiga, A. Preliminary Study on Pasta Samples Characterized in Antioxidant Compounds and Their Biological Activity on Kidney Cells. *Nutrients 2021*, 13, 1131.

[10] Koch, W. Dietary Polyphenols-Important Non-Nutrients in the Prevention of Chronic Noncommunicable Diseases. A Systematic Review. Nutrients 2019, 11, 39.

[11] Chen, Q.; Espey, M.G.; Krishna, M.C.; Mitchell, J.B.; Corpe, C.P.; Buettner, G.R.; Shacter, E.; Levine, M. Pharmacologic ascorbic acid concentrations selectively kill cancer cells: Action as a pro-drug to deliver hydrogen peroxide to tissues. *Proc. Natl. Acad. Sci.* USA 2005, 102, 13604–13609

[12] Noce, A.; Marrone, G.; Di Daniele, F.; Di Lauro, M.; Pietroboni Zaitseva, A.; Wilson Jones, G.; De Lorenzo, A.; Di Daniele, N. Potential Cardiovascular and Metabolic Beneficial Effects of omega-3 PUFA in Male Obesity Secondary Hypogonadism Syndrome. *Nutrients 2020*, 12, 2519.

[13] Owen, A.J.; Abramson, M.J.; Ikin, J.F.; McCaffrey, T.A.; Pomeroy, S.; Borg, B.M.; Gao, C.X.; Brown, D.; Liew, D. Recommended Intake of Key Food Groups and Cardiovascular Risk Factors in Australian Older, Rural-Dwelling Adults. *Nutrients 2020*, 12, 860.

[14] Kee, S. K., Son, Y. J, Kim H.G., Lee J. Il., Cho, H.S., Lee, S., 2014, Associations between food and bever-age groups and major diet-related chronic diseases: an exhaustive review of pooled/meta-analyses and systematic reviews, *Nutr Rev. 2014 Dec*; 72(12):741-62. doi: 10.1111/nure.12153

[15] Qudsi, D., Kartiwi, M., Saleh, N.B., 2017, Predictive data mining of chronic diseases using decision tree: A case study of health insurance company in Indonesia. *International Journal of Applied Engineer-ing Research 12(7)*:1334-1339

[16] Lei Z., Yang, S., Liu, H., Aslam, S., Liu, J., Bugingo, E., Zhang, D., 2018, Mining of Nutritional Ingredi-ents in Food for Disease Analysis, *IEEE Access 6(1)*:52766-52778

[17] McCabe, R.M, Adomavicius, G., Johnson P.E., Rund, E., Rush, A., Sperl-Hillen, A., 2008, Using Data Mining to Predict Errors in Chronic Disease Care, Advances in Patient Safety, *New Directions and Alternative Approaches in Vol. 3*: Performance and Tools.

[18] Haslam, D.W., James, W.P.T., Obesity, In the Lancet, Volume 366, Issue 9492, Pages 1197-1209

[19] Lange, K.W., James W.P.T., Makulska-Gertruda E., Nakamura Y., Reissmann, A., 2008, A. Sperl-Hillen, Using Data Mining to Predict Errors in Chronic Disease Care, Advances in Patient Safety. *New Directions and Alternative Approaches* (Vol. 3: Performance and Tools)

[20] Hearty, A.P., Gibney, M.J., 2008, A. Richonnet, C., Mazur, A., Analysis of meal patterns with the use of supervised data mining techniques—artificial neural networks and decision trees, *American Journal of Clinical Nutrition*, Volume 88, Issue 6, Pages 1632–1642.

[21] Von Ruesten, A., Feller, S., Bergmann, N.M., Boeing, H., 2013, S., Diet and risk of chronic diseases: results from the first 8 years of follow-up in the EPIC-Potsdam study, *European Journal of Clinical Nutrition volume 67*, pages412–419.

[22] Yu E. Y. W., Wesselius A., Sinhart C., Wolk A., 2020, A data mining approach to investigate food groups related to incidence of bladder cancer, *Bladder cancer Epidemiology and Nutritional Determinants International Study*, Cambridge University Press

[23] Einsele, F., Sadeghi, L., Ingold, R., Jenzer, H., 2015, A Study about Discovery of Critical Food Consumption Patterns Linked with Lifestyle Diseases using Data Mining Methods, *HealthInf, BIOSTEC - International Joint Conference on Biomedical Engineering Systems and Technologie*s, Lisbon.

[24] Mewes I., Jenzer H., Einsele, F., 2021, A Study about Discovery of Critical Food Consumption Patterns Linked with Lifestyle Diseases for Swiss Population using Data Mining Methods, *14th International Conference on Health Informatics*

[25] Lustenberger T., Jenzer H., Einsele F., 2022, Discovery of Association Rules of the Relationship between Food Consumption and Lifestyle Diseases from Swiss Nutrition's (MENUCH) Dataset & Multiple Swiss Health Datasets from 1992 To 2012, *6th International Conference on Big Data & Health BDHI*, AISCA, NET, DNLP, BDHI – 2022, pp. 77-93, 2022

AUTHORS

Farshideh Einsele, Prof. Dr., Lecturer & researcher in the Business section of BUAS, she teaches business informatic subjects and her research work is dedicated to epidemiology, big data and data mining

Jonas Baschung, Student in Bern University of Applied Sciences. He currently fulfilled his Bachelor in business information systems.