

Mild Cognitive Impairment Detection Using Association Rules Mining

František Babič¹ , Ľudmila Puzstová¹, Ljiljana Trtica Majnarić^{2,3} 

Abstract

A single Mild cognitive impairment (MCI) is a transitional state between normal cognition and dementia. The typical diagnostic procedure relies on neuropsychological testing, which is insufficiently accurate and does not provide information on patients' clinical profiles. The objective of this paper is to improve the recognition of elderly primary care patients with MCI by using an approach typically applied in the market basket analysis – association rules mining. In our case, the association rules represent various combinations of the clinical features or patterns associated with MCI. The analytical process was performed in line with the CRISP-DM, the methodology for data mining projects widely used in various research or industry domains. In the data preparation phase, we applied several approaches to improve the data quality like the k-Nearest Neighbour, correlation analysis, Chi Merge and K-Means algorithms. The analytical solution's success was confirmed not only by the novelty and correctness of new knowledge, but also by the form of visualization that is easily understandable for domain experts. This iterative approach provides a set of rules (patterns) that meet minimum support and reliability. The extracted rules may help medical professionals recognize clinical patterns; however, the final decision depends on the expert. A medical expert has a crucial role in this process by enabling the link between the information contained in the rules and the evidence-based knowledge. It markedly contributes to the interpretability of the results.

Keywords: Association rules, Patterns, Mild cognitive impairment, Interpretability.

1 Introduction

Mild cognitive impairment (MCI) is defined as a grade of cognitive impairment that is severe enough to be noticed by other persons and to register on cognitive tests but insufficiently strong to interfere with the activities of daily life (Albert et al., 2011). MCI was coined to emphasize the gradually changing continuum along the course of cognitive function decline, from normal cognition to dementia. This concept is expected to have significant practical benefits. The sooner the cognitive disorder is diagnosed, and the treatment starts, the more

¹ Department of Cybernetics and Artificial Intelligence, Faculty of Electrical Engineering and Informatics, Technical University of Košice, Letná 9, 042 00 Košice, Slovak Republic

✉ ľudmila.puzstova.2@tuke.sk

² Department of Public health, Faculty of Dental Medicine and Health, Josip Juraj Strossmayer University of Osijek, Crkvena 21, 31000 Osijek, Croatia

³ Department of Internal medicine, Family Medicine and the History of Medicine, Faculty of Medicine, Josip Juraj Strossmayer University of Osijek, Josipa Huttlera 4, 31000 Osijek, Croatia

likely the progression of memory loss is prevented or delayed (Lazarczyk et al., 2012). However, this concept is not easy to apply in everyday practice. The main concerns are how to recognize high-risk patients for cognitive impairment and how to identify those of them who will likely progress from MCI to dementia. MCI is a widespread disorder in the elderly population which affects 10–20% of people aged 65 or more (Albert et al., 2011). It can be recognized by using neuropsychological tests, of which the Mini-Mental State Examination (MMSE) is the most widely used in elderly primary care (PC). However, only a part of the positively tested people develops clinically overt dementia. Yet, dementia is an emerging diagnosis in modern societies and is more and more considered a public health problem. It is because it is associated with a rapid loss of individuals' functional independence and the need for long-term nursing home care, which poses a considerable burden on the wealth of their families and a financial burden on healthcare systems (Albert et al., 2011). Moreover, dementia is considered a member of the group of common aging diseases, which also includes hypertension, diabetes, cardiovascular (CV) disease and some cancers, which are known to share the common pathophysiology background and are clinically presented with overlapping comorbidities (Buchanan, 2006).

Despite the increasing awareness about these issues, the current data modelling procedures are still focused on improving the diagnostic accuracy of the MMSE and other cognitive tests. For this purpose, data from magnetic resonance imaging or other imaging techniques are usually added to the model, while information on patient sociodemographic and clinical features is rarely included (Qiu et al., 2018; Zhang et al., 2014). One of the reasons is that a methodological framework which would be able to capture the complexity of the clinical phenotypes is yet insufficiently developed.

We aimed to identify clinical features and patterns for patients with MCI, which is assumed to improve the recognition of patients with MCI. This would be of great practical importance for PC providers who encounter most elderly people in the population and perform the screening procedure for MCI. The MMSE, if performed at population level, is time-consuming and requires high patient engagement. Also, the ability of the test to accurately distinguish between those with normal cognition and MCI is not sufficiently high (Aevarsson & Skoog, 2000). Also, knowing the clinical features of persons with MCI may indicate the pathogenetic pathways associated with this disorder, which could guide research on the factors and mechanisms of the progression from MCI to dementia. We decided to evaluate experimentally the potential of association rules mining for this task.

The whole paper is organized as follows: a short introduction of the disease with motivation and used methodology. In the related work, we present selected existing studies to support our decision to apply the association rules mining for this type of task and to eliminate known bottlenecks like the first big number of extracted rules by domain knowledge provided by cooperating experts. The performed diagnostic process is described following the CRISP-DM methodology. The conclusion summarizes our results and their usability in primary care.

1.1 CRISP-DM methodology

The analytical process was conducted in close cooperation between medical doctors, PC providers, and data analysts. Each of these roles has its knowledge, experience, and skills, e.g. medical experts provide a relevant interdisciplinary context for heterogeneous data and possible positive diagnoses. Data analytics is an iterative and interactive process that brings new, potentially useful knowledge. It is essential to define a common vocabulary and a framework for cooperation. For this purpose and based on our expertise from different domains, we decided to use the Cross-industry standard process for data mining methodology

typically used in the field of data analytics (Chapman et al., 2000; Shearer, 2000). This methodology defines six main phases, specifically business understanding, data understanding, data preparation, modelling, evaluation, deployment (Fig. 1).

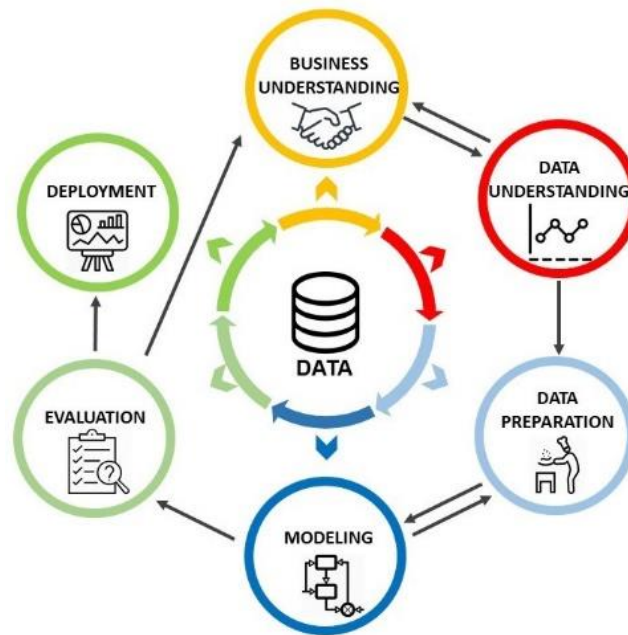


Fig. 1. CRISP-DM process model. Source: Authors.

Business understanding deals with a specification of business goals followed by transformation to a specific analytical task. Data understanding starts with a collection of necessary data for the specified task and ends with a detailed description, including some statistical characteristics. Data preparation is usually the most complex and the most time-consuming phase. Generally, it takes 60-75% of the overall time. It contains data aggregation (different data samples), cleaning (missing or incorrect values), reduction (similar attributes, hidden relationships, irrelevant attributes), or transformation (derived attributes, discretization, normalization). Modelling deals with an application of suitable machine learning methods on the pre-processed data. Also, in this step it is necessary to specify the correct metrics for results evaluation, e.g. accuracy, ROC, precision, recall, etc. The evaluation phase is oriented towards the evaluation of generated models and obtained results based on specified goals in business understanding. This phase requires intensive cooperation between data analysts and domain experts. The deployment consists of the exploration of created models in real cases, their adaption, maintenance, and collection of acquired experience and knowledge.

1.2 Association rules mining

The typical representation of the association rule is $X \Rightarrow Y$, where X, Y are subsets from a whole set of items I . The quality of each generated rule is evaluated by two metrics: support and confidence. Support is an indication of how frequently the itemset appears in the dataset. Confidence is an indication of how often the rule was found to be true. The values of these metrics are set when the algorithm Apriori starts to reduce the search space. The Apriori algorithm was designed by R. Agrawal and R. Srikant in 1994 (Agrawal & Srikant, 1994) to find frequent itemsets from data. The algorithm is based on finding frequent itemsets, which represents combinations/conjunctions of the attribute category meeting the minimum support value. The basic approach uses the breadth-first search strategy (a strategy of searching for the

shortest path between two nodes in the graph structure) based on measuring the support (similarity) and confidence (dissimilarity) between sets of items (Fig. 2).

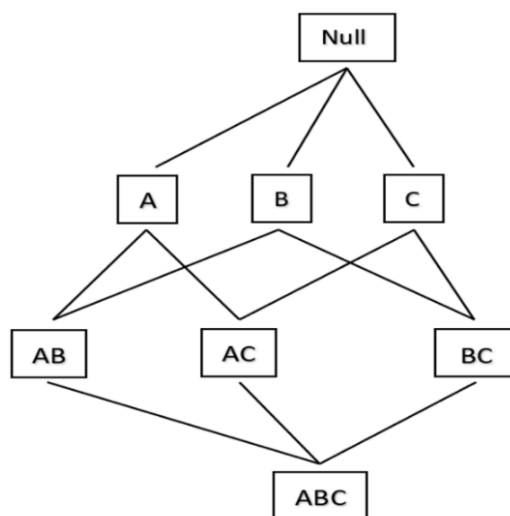


Fig. 2. The lattice of itemset. Source: (Agrawal & Srikant, 1994).

1.3 Related work

Association rule mining is a popular data mining technique because of its easily interpretable results, and it is used in many research domains. Its typical application is market basket analysis representing one of retailers' essential techniques to identify an association between offered items. We are convinced that this approach can also be successful in the medical diagnosis process. Several existing works confirm this assumption. For example, Sariyer and Tasar (2019) used the Apriori algorithm to extract hidden patterns and relations between diagnosis and diagnostic test requirements in medical data received from an emergency department. The diagnoses were grouped into 21 categories based on the ICD standard. The laboratory tests were grouped into four main categories like hemogram, biochemistry, cardiac enzyme, urine, and human excrement related. An expert evaluated all the extracted rules.

The authors concluded that understanding the association between a patient's diagnosis and diagnostic test requirements can improve decision-making and be used to support physicians. Alwidian et al. (2018) used weighted classification based on association rules algorithm for predicting breast cancer. The authors also applied a new pruning and prediction technique based on statistical measures to generate more accurate rules. Their model outperformed other association classification algorithms like CBA, CMAR, or FACA. Borah and Nath (2018) used association rules mining to generate a new set of rare association rules from updated medical databases to identify the symptoms and risk factors for three adverse diseases: cardiovascular disease, hepatitis, and breast cancer. They focused on the notion that all necessary data for association rules mining must be available at the beginning. Their algorithm can insert or delete a case online without re-executing the entire mining process.

Harahap et al. (2018) focused on the importance of an appropriate selection process of required medicine based on the development of the patient's illness. They analyzed patient prescriptions to identify the relationship between the disease and the physician's medicine in treating the patient's illness. Firstly, they used the k-means algorithm for clustering to 10 diseases and then applied the Apriori algorithm for finding association rules based on support, confidence, and lift value. For grouping, they used three variables – age, gender, and disease. The highest number of instances was in cluster 0 – unspecified cataract disease. Between the

top ten diseases (antecedent) and the related medicine (consequent) was the minimum limit value of support 20% and confidence 65%. Lakshmi and Vadivu (2017) extracted association rules from medical health records using multi-criteria decision analysis. They used a lift value to select interesting rules for the next clinical validation. In general, the authors consider this method useful for identifying precise clinical associations between medications, laboratory results, and comorbidities.

2 Diagnostic Process

The following chapters describe the respective phases of CRISP-DM performed by the international research team.

2.1 Business understanding

The typical diagnostic procedure of Mild cognitive impairment relies on neuropsychological testing, which is insufficiently accurate and does not provide information on patient clinical profiles. Knowing the clinical features of persons with MCI may indicate the pathogenetic pathways associated with this disorder, which could guide research on the factors and mechanisms of progression from MCI to dementia. From the business point of view, it is interesting to have these pathways supporting the decision process of medical experts. From an analytical point of view, we aimed to provide a set of diagnostic rules in a simple, understandable form to meet this business goal. In parallel, we tested and evaluated the potential of association rules mining for this type of task. This decision also covered the challenge of how to reduce the first large volume of extracted association rules. For this purpose, we applied the domain knowledge provided by an intensive collaboration framework between medical experts, data analysts, and primary care providers.

2.2 Data understanding

The study included 93 participants, 35 M/58 F, 47–89 years old (median 69 years). They were recruited from several general medical practices in the town of Osijek (about 80,000 inhabitants) in eastern Croatia, a region with high rates of CV disease. Only participants who gave their signed, informed consent were included in the study. Data collection for one person did not last longer than six months. The data for this analysis were obtained from a multicomponent dataset specifically established for practicing data mining methods. The data were collected over three years (2007–2010).

The study was conducted following with the Declaration of Helsinki and approved by the Ethics Committee of the Faculty of Medicine, University of Zagreb (04-76/2006-396).

The data were low-cost, easily available parameters collected to determine the health status of the examined patients (Table 1). A large proportion of these parameters is routinely collected in PC electronic health records (eHRs). For example, nominal parameters that are known to influence the level of inflammation and circulation properties like age and gender, diagnoses of the main groups of chronic diseases, and continuous use of some medications, including statins (hypolipidemic agents), analgesics and anticoagulant/antiaggregant drugs (Antonopoulos et al., 2012). Information on adverse drug reactions was used to indicate inappropriate polypharmacy prescriptions, which is usually the case in multimorbid and frail elderly persons (Tjia et al., 2010).

Abbreviation	Parameter description	Abbreviation	Parameter description
age	Age (years)	CRP	C-reactive protein (mg/L)
sex	M=Male, F=Female	E	Erythrocytes number x1012/L
Hyper	Hypertension (yes, no)	HB	Hemoglobin (g/L)
DM	Diabetes mellitus (yes, IGT=Impaired glucose tolerance, No)	HTC	Hematocrit (erythrocyte volume blood fraction)
F Glu	Fasting blood glucose (mmol/L)	MCV	Mean cell Volume (fL)
HbA1c	Glycosylated Hemoglobin (%) - showing average blood glucose during last three months	FE	Iron (g/L)
Chol	Total Cholesterol (mmol/L)	PROT	Total serum proteins (g/L)
TG	Triglycerides (mmol/L)	ALB	Serum albumin (g/L)
HDL	HDL-cholesterol (mmol/L)	clear	Creatinine clearance (ml/s/1.73m2)
Statins	Therapy with statins (yes, no)	HOMCIS	Homocysteine (μmol/L)
Anticoag	Therapy with anticoagulant/antiaggregant drugs (yes, no)	ALFA1	Serum protein electrophoresis (g/L)
CVD	Cardiovascular diseases as myocardial infarction, angina, history of revascularization, stroke, transient ischemic cerebral event, peripheral vascular disease (yes, no)	ALFA2	Serum protein electrophoresis (g/L)
BMI	Body Mass Index (kg/m2)	BETA	Serum protein electrophoresis (g/L)
w/h	Waist/hip ratio	GAMA	Serum protein electrophoresis (g/L)
Arm cir	Mid arm circumference (mm)	RF	Rheumatoid Factor level (IU/ml)
skinf	Triceps skinfold thickness (mm)	VITB12	Vitamin B12 (pmol/L)
gastro	Gastroduodenal disorders as gastritis, ulcer (yes, no)	FOLNA	Folic acid (mM/L)
Uro	Chronic urinary tract disorders (yes, no) - recurrent cystitis in women, symptoms of prostates in men	INS	Insulin (μIU/L)
COPB	Chronic obstructive pulmonary disease (yes, no)	TSH	Thyroid-stimulating hormone (IU/ml)
Aller d	Allergy (Rhinitis and/or Asthma) (yes, no)	FT4	Free thyroxine (pmol/L)
dr aller	Drugs allergy (yes, no)	FT3	Free triiodothyronine (pmol/L)
analg	Therapy with analgesics/NSAR (yes, no)	ANA	Antinuclear antibodies (autoantibodies) (μIU/ml)
derm	Chronic skin disorders as chronic dermatitis, dermatomycosis (yes, no)	IGE	IgE (kIU/L)

Neo	Malignancy (yes, no)	LE	Leukocytes Number x109/L
OSP	Osteoporosis (yes, no)	Psy	Neuropsychiatric disorders as anxiety/depression, Parkinson`s disease, cognitive impairments (yes, no)
CMV	Cytomegalovirus specific IgG antibodies (IU/ml)	HBG	Helicobacter pylori specific IgG (IU/ml)
HPA	Helicobacter pylori specific IgA (IU/ml)	MMSE	Mini Mental State Examination

Tab. 1. Description of parameters. Source: Authors.

Anthropometric measurements, if not updated in eHRs, were utilized at patient encounters (Avila-Funes et al., 2009; Whitmer, 2007). Several laboratory tests were chosen to indicate age-related pathophysiologic changes, including information on the level of inflammation, nutritional status, chronic renal impairment, CV metabolic factors, and thyroid gland hormones, which are all widely cited health-related risk factors for cognitive impairment (Bugnicourt et al., 2013; Hogervorst et al., 2010; Postiglione et al., 2001; Roberts et al., 2009; Umegaki, 2014). The results of some laboratory tests were used for this, as these are part of regular chronic disease surveillance programs. For more specific biochemical and hematological tests, patients were referred to the central laboratory of the Osijek Clinical Hospital for a venipuncture. All laboratory tests were performed according to the standard procedures. The medical experts used creatinine clearance and serum homocysteine as measures of renal function decline. Increased serum homocysteine concentrations (hyperhomocysteinemia) were reported as a CV risk factor (Postiglione et al., 2001). Insulin measurements in a fasting state were obtained to approximate the level of insulin resistance (Schrijvers et al., 2010). The degree of inflammation was indicated by C-reactive protein (CRP), total leukocyte count, and serum protein electrophoresis fractions (Roberts et al., 2009; Jain et al., 2011). We also used other laboratory parameters to enlarge the scope of possible factors for the association rules, including parameters indicating common chronic latent infections and disturbed age-related immune reactions (Cavagna et al., 2012; Deleidi et al., 2015; Futagami et al., 1998).

The participants were also screened for cognitive impairment using the MMSE, the most widely used screening test for assessing cognitive function, validated in many populations, including the Croatian elderly population (Boban et al., 2012). A score of 24 or less (out of the maximum of 30) indicates cognitive impairment. The test is relatively sensitive in diagnosing overt dementia but is less accurate in distinguishing cognitively healthy individuals from those with MCI.

2.3 Data preparation

At first, we performed an exploratory data analysis to visually clarify the relationships between some input parameters and MCI diagnosis as the output parameter (Fig. 3). As an example, we provide this figure illustrating the fact that almost all patients with MCI are 61 years or older. The generated graphs were consulted with the domain expert to detect a possible outlier or important knowledge for the modelling phase.

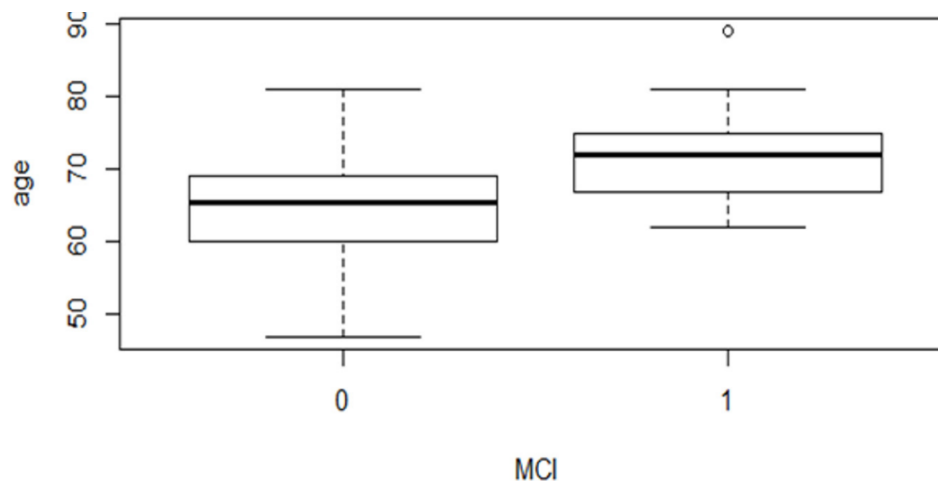


Fig. 1. Boxplot representing the relationships between the input parameter of age (y-axis) and the target parameter of MCI (x-axis). Source: Authors.

The dataset contained 1.64% missing values replaced by the K-Nearest Neighbours (k-NN) algorithm (Altman, 1992) to improve data quality and provide the conditions for various machine learning methods. The k-NN provides a point with its closest k neighbours in a multi-dimensional space. The missing values were approximated by the values of the points (patients) closest to it based on the other available input variables. The initial value of the k parameter was experimentally evaluated.

Next, we investigated possible confounding relationships between input variables using the correlation analysis. Based on the minimal threshold of 0.8 for strong correlations, we identified two pairs of intercorrelated parameters: (HB) vs. (HTC) (0.963) and (E) vs. (HTC) (0.812).

Association rules mining, including the Apriori algorithm, requires a discretisation of the numerical attributes. Thus, the parameter of body mass index (BMI) was transformed according to the standard categorization of the World Health Organization, cited elsewhere, where $BMI < 18.5$ indicates underweight, $18.5 < BMI < 24.99$ indicates normal weight, $BMI \geq 25$ indicates overweight and $BMI \geq 30$ indicates obesity. The relationships between the input parameter BMI and the target parameter MCI are depicted in Fig. 4. The median values of the parameter MCI are different in obese and overweight persons (above the cut-off score for MCI diagnosis) compared to persons with normal weight (Mann-Whitney-Wilcoxon test, $p=0.07$).

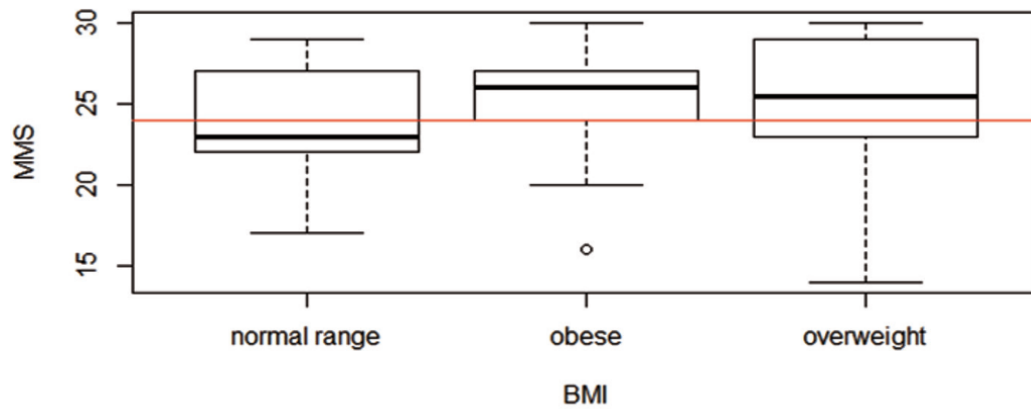


Fig. 2. Boxplot representing the relationships between the input parameter of BMI (x-axis) and the target parameter of MCI (y-axis). Source: Authors.

We transformed other numerical variables within two methods (Tab. 2): unsupervised K-Means clustering and supervised Chi Merge. These methods were used instead of the typical discretization procedure with a fixed-length window. Some relevant components may be lost if cut-offs are incorrectly placed within the pattern.

The Chi Merge algorithm uses the χ^2 statistic to discretize numeric parameters (Kerber, 1992). It accounts for significant numbers of categories (segments) when constructing discretization intervals. Because the Chi Merge algorithm tends to construct many categories, as an alternative discretization method, we used the K-Means clustering algorithm (Lloyd, 1982).

Attribute	ChiMerge	K-means
age	(0, 61.5>, (61.5, 70.5>, (70.5, inf)	<47, 63.2), <63.2, 73.2), <73.2, 89>
FGlu	(0, 7.05>, (7.05, inf)	<4.6, 5.97), <5.97, 8.93), <8.93, 13.90>
HbA1c	(0, 3.505>, (3.505, 3.615>, (3.615, 3.645>, (3.645, 3.825>, (3.825, 3.845>, (3.845, 6.685>, (6.685, 8.885>, (8.885, inf)	<2.89,5.90), <5.90, 28.00), <28.00, 48.30>
Chol	(0,4.55>, (4.55, 6.55>, (6.55, 7.05>, (7.05, 8.65>, (8.65, inf)	<3.00, 5.56), <5.56, 7.23), <7.23, 8.90>
TG	(0, 0.55>, (0.55, 1.05>, (1.05, 1.65>, (1.65, 2.15>, (2.15, inf)	<0.5, 1.79), <1.79, 4.87), <4.87, 9.30>
HDL	(0, 0.945>, (0.945, 1.190>, (1.190, 1.380>, (1.380, 1.525>, (1.525, 1.565>, (1.565, 1.595>, (1.595, 1.745>, (1.745, 1.855>, (1.855, inf)	<0.87, 1.38), <1.38, 1.89), <1.89, 2.53>
skinf	(0, 26.5>, (26.5, 28.5>, (28.5, inf)	<16.0, 27.4), <27.4, 36.5), <36.5, 50.0>
CRP	(0, 4.25>, (4.25, 4.95>, (4.95, 5.15>, (5.15, 5.25>, (5.25, 6.45>, (6.45, 7>, (7, 13.05>, (13.05, inf)	<0.80, 6.48), <6.48, 14.9), <14.9, 24.5>
CLEAR	(0, 2.055>, (2.055, 2.155>, (2.155, 2.315>, (2.315, inf)	<0.72, 1.36), <1.36, 1.88), <1.88, 3.21>
HOMCIS	(0, 6.1>, (6.1, 8.25>, (8.25,12.15>, (12.15, 15.25>, (15.25, 16.45>, (16.45, 18.6>, (18.6, 19.9>, (19.9, inf)	<5.0, 11.3), <11.3, 16.5), <16.5, 25.9>

	inf)	
INS	(0, 14.65>, (14.65, 15.15>, (15.15, inf)	<5.9, 26.2), <26.2, 92.8), <92.8, 149.7>

Tab. 2. Examples of discretized parameters with specified cut-off values. Source: Authors.

For example, the parameter fasting glucose (FGlu) was divided into three ranges that are very similar to the standard distributions of blood glucose levels used in practice (Umegaki, 2014). In general, K-Means discretisation method provided the result closer to the typical cut-off values described in the literature. Modelling phase was not customized to these discretisation methods; the aim was to provide a broader view of target diagnostics.

2.4 Modelling

We applied the Apriori version available within R language (package “arules”) to the several data samples prepared in the pre-processing phase as the subset containing only CV risk factors. As CV risk factors, we considered the following parameters: age, sex, Hyper, DM, Fglu, HbA1c, Chol, TG, HDL, statins, CVD, BMI, w/h, Arm cir, skinf, Anticoag, Clear, INS, HOMCIS, CRP.

We performed several iterations, and the experiments brought a big set of rules. The first filter was applied based on the minimal values of support and confidence. For example, this step resulted in 25, 10, 30, and 55 rules. Simple visualization inspired by the tree structure was provided to the medical expert for her evaluation: new or interesting knowledge, more complex rather than elementary, without redundancy (Tables 3–6). They can be used in two ways:

- in clinical practice to help medical doctors select older people with MCI,
- to clarify the clinical context associated with MCI.

No	Rule (antecedent)	supp	conf
1	age=(70,5;inf),TG=(1,05;1,65>,Statins=No,Clear=(0;2,055>	0.12	1.00
2	age=(70,5;inf),sex=F,Clear=(0;2,055>,INS=(15,15;inf>	0.12	0.92
3	age=(70,5;inf),sex=F,CVD=No,Clear=(0;2,055>	0.12	0.92
4	Fglu=(0;7,05>,DM=No,Statins=No,CRP=(4,25;4,95>,Clear=(0;2,055>	0.12	0.92
5	Fglu=(0;7,05>,Hyper=Yes,DM=No,CRP=(4,25;4,95>,Clear=(0;2,055>	0.11	0.91
6	age=(70,5;inf),sex=F,skinf=(28,5;inf),Clear=(0;2,055>,INS=(15,15;inf>	0.11	0.91
7	age=(70,5;inf),sex=F,Statins=No,Clear=(0;2,055>,INS=(15,15;inf>	0.11	0.91
8	age=(70,5;inf),CVD=No,skinf=(28,5;inf),Clear=(0;2,055>,INS=(15,15;inf>	0.11	0.91
9	age=(70,5;inf),Statins=No,CVD=No,skinf=(28,5;inf),Clear=(0;2,055>	0.12	0.92
10	age=(70,5;inf),Hyper=Yes,CVD=No,skinf=(28,5;inf),Clear=(0;2,055>	0.11	0.91
11	age=(70,5;inf),Hyper=Yes,CVD=No,skinf=(28,5;inf),Clear=(0;2,055>	0.11	0.91

Tab. 3. Extracted rules by the Apriori algorithm for the subset of data with only CV risk factors included (min. support=0.11; min. confidence=0.85, Chi Merge discretization, meaning, MCI=yes). Source: Authors.

Interpretation of rule no. 1: IF the age of the patient is greater than 70.5, triglycerides between 1.05 and 1.65, without therapy with statins and creatinine clearance between 0 and 2.055

THEN this patient is MCI positive with 0.12 support and 100% confidence. It means that the whole dataset contains 10 cases (12%) that meet the conditions of this rule, and all these cases (100%) are positive for MCI diagnosis.

No	Rule (antecedent)	supp	conf
1	age=<63,2,73,2),TG=<0,50,1,79),CRP=< 0,80, 6,48),Clear=<1,36,1,88)	0.09	0.89
2	age=<63,2,73,2),TG=<0,50,1,79),HOMCIS=<11,3,16,5),INS=<5,9, 26,2)	0.09	0.89
3	age=<63,2,73,2),HbA1c=< 2,89, 5,90),TG=<0,50,1,79),HOMCIS=<11,3,16,5)	0.09	0.89
4	Statins= No, BMI=overweight, TG=<0,50,1,79), HOMCIS=<11,3,16,5), INS=<5,9, 26,2)	0.09	1.00
5	Statins= No,age=<63,2,73,2),TG=<0,50,1,79),CRP=< 0,80, 6,48) ,Clear=<1,36,1,88)	0.09	0.89
6	age=<63,2,73,2),HbA1c=< 2,89, 5,90), TG=<0,50,1,79), HOMCIS=<11,3,16,5),INS=<5,9, 26,2)	0.09	0.89

Tab. 4. Extracted rules by the Apriori algorithm for the subset of data with only CV risk factors included (min. support=0.08; min. confidence=0.85, k-means discretization, meaning, MCI=yes). Source: Authors.

No	Rule (antecedent)	supp	conf
1	age=(70,5;inf),COPB=No,dr_aller=No,GAMA=(0;15,9>,Clear=(0;2,055>,FT4=(12,65;inf)	0.16	0.94
2	age=(70,5;inf),aller_d=No,neo=No,GAMA=(0;15,9>,Clear=(0;2,055>,FT4=(12,65;inf)	0.17	1.00
3	age=(70,5;inf),aller_d=No,dr_aller=No,neo=No,GAMA=(0;15,9>,FT4=(12,65;inf)	0.16	1.00
4	age=(70,5;inf),COPB=No,dr_aller=No,HTC=(0,355;inf),GAMA=(0;15,9>,FT4=(12,65;inf)	0.16	0.94
5	age=(70,5;inf),COPB=No,dr_aller=No,HTC=(0,355;inf),GAMA=(0;15,9>,Clear=(0;2,055>,FT4=(12,65;inf)	0.16	0.94
6	age=(70,5;inf),COPB=No,aller_d=No,dr_aller=No,HB=(116,5;inf),GAMA=(0;15,9>,Clear=(0;2,055>, FT4=(12,65;inf)	0.16	0.94
7	age=(70,5;inf),COPB=No,dr_aller=No,HB=(116,5;inf),HTC=(0,355;inf),GAMA=(0;15,9>,Clear=(0;2,055>, FT4=(12,65;inf)	0.16	0.94
8	age=(70,5;inf),COPB=No,aller_d=No,dr_aller=No,HB=(116,5;inf),HTC=(0,355;inf), GAMA=(0;15,9>, FT4=(12,65;inf)	0.16	0.94
9	age=(70,5;inf),COPB=No,aller_d=No,dr_aller=No,HB=(116,5;inf),HTC=(0,355;inf), GAMA=(0;15,9>, Clear=(0;2,055>,FT4=(12,65;inf)	0.16	0.94

Tab. 5. Extracted rules by the Apriori algorithm for the whole dataset (min. support=0.16; min. confidence=0.9, Chi Merge discretization, meaning, MCI=yes). Source: Authors.

No	Rule (antecedent)	supp	conf
1	Gastro= No,aller_d= No,neo= No,age=<73,2,89,0>,RF=<9,0, 47,8)	0.12	0.92
2	aller_d= No,dr_aller= No,neo= No,age=<73,2,89,0>,RF=<9,0, 47,8),IGE=<2, 262)	0.12	0.92
3	COPB= No,aller_d= No,age=<63,2,73,2),TG=<0,50,1,79),HPA=<2,3,	0.12	0.92

	47,6),INS=<5,9, 26,2)		
4	aller_d= No,age=<63,2,73,2),TG=<0,50,1,79),HPA=<2,3, 47,6),RF=<9,0, 47,8),ANA=<7, 33)	0.12	0.92
5	sex=F,Hyper=Yes,COPB= No,aller_d= No,HPA=<2,3, 47,6),MCV=< 87,3, 93,7),RF=< 9,0, 47,8),ANA=<7, 33)	0.12	0.92
6	Hyper=Yes,aller_d= No,HPA=<2,3, 47,6),CRP=< 0,80, 6,48),MCV=< 87,3, 93,7),FE=<13,2,19,4),RF=< 9,0, 47,8),ANA=<7, 33),IGE=<2, 262)	0.12	0.92
7	sex=F,COPB= No,aller_d= No,dr_aller= No,neo= No,FE=<13,2,19,4),RF=< 9,0, 47,8),ANA=<7, 33),IGE=< 2, 262)	0.12	0.92
8	sex=F,Hyper=Yes,COPB= No,aller_d= No,neo= No,HPA=<2,3, 47,6),MCV=<87,3, 93,7),RF=< 9,0, 47,8),ANA=<7, 33)	0.12	0.92

Tab. 6. Extracted rules by the Apriori algorithm for the whole dataset (min. support=0.11; min. confidence=0.9, k-means discretization, meaning, MCI=yes). Source: Authors.

2.5. Evaluation

From the data mining point of view, the crucial aspect was to reduce effectively the initial set of generated rules and provide a comprehensive overview for the domain expert. The expert provided a deeper analysis of these results and we present only part of it as an example.

A high proportion of examined patients, 37 out of 93 (39.8%), were found to be diagnosed with MCI. This proportion is higher than it is usually reported in epidemiologic studies and can be explained by a high burden of chronic diseases due to recent negative socioeconomic trends in the region. As it is elsewhere reported for cognitive disorders, women dominated over men (F/M = 25/12), and they mostly belonged to the elderly population group (67–75 years).

The conclusion which arises from these results is that patients with MCI show low-level variation and they all belong to the standard clinical framework. For example, the extracted rules shared the following six parameters: statins, age, TG, INS, CRP, and Clear, indicating CV risk factors (Tables 3 and 4). These parameters indicate the important mechanisms of ageing diseases, such as increased insulin resistance (parameters INS and TG), a low-level chronic inflammation (parameters statins and CRP), and chronic renal impairment (parameter Clear), for which evidence also suggests their involvement in the development of cognitive impairment (Antonopoulos et al., 2012; Bugnicourt et al., 2013; Roberts et al., 2009; Schrijvers et al., 2010).

The aging, chronic conditions share a common background and pathophysiology pathways (Buchanan, 2006). Therefore, if the input dataset is sufficiently large, many of the parameters extracted in the rules would be complementary to each other. In other words, these parameters cluster together, indicating the common pathways. The knowledge needed to understand the complementarity of CV risk factors is exacting and requires the active input of a medical expert.

The clinical phenotype of MCI, indicated by the rules that are presented in Table 3, is a characteristic of women, aged 70.5 or more (as shown by the parameters: age=70.5 and sex=F). The clinical features of this phenotype may be reconstructed from the interval-values of the parameters that are most frequently present in these rules, including INS, TG, Clear, and skinf (Table 3). This syndrome is characterized by increased inflammation and muscle wasting, the condition called frailty (in the rules indicated by higher interval-values of the

parameters skinf and CRP). Also, we identified disturbed glucose-related metabolism, called insulin resistance, which in this syndrome is specifically presented with high serum insulin values and low values of the CV risk factors, fasting serum glucose and serum triglycerides (Walker et al., 2017; Whaley-Connell & Sowers, 2018).

In the rules related to CV risk factors performed by another discretization method (Table 4), the parameter Clear, an indicator of low renal function, is substituted with its complementary parameter HOMCIS. This alternative indicates variations in serum concentrations of the substance homocysteine and is a marker of impaired renal function (Van Guldener, 2006). This difference in the composition of the rules presented in Tables 3 and 4, concerning the alternate presence of the complementary parameters Clear and HOMCIS, may be the additional expression of the same chronic renal impairment syndrome. Another possibility is that these groups of rules indicate two variants of this syndrome, which differ from each other in the levels of serum homocysteine concentrations (Lloyd, 1982).

When we used the whole dataset, the rules (Tables 5 and 6) contained some new parameters, besides those indicating CV risk factors, indicating the even broader clinical context of MCI, and posing new hypotheses. These hypotheses must be confirmed in future studies. Many rules consisted of a limited number of uniformed phrases, indicating several independent functional units, or pathways, which operate within the pathophysiology framework of MCI.

2.6. Deployment

The cooperation between primary care providers, medical experts and data analysts resulted in a new set of knowledge confirmed in relevant conditions and based on the respective data sample. It represents only the first step in the verification and approval process, but it is essential to make these first steps to generate diagnostic prototype or, in other words, Clinical Decision Support System.

3 Discussion and Conclusions

The Association rule mining is a useful method for mapping potentially relevant parameters in multicomponent datasets, in problem-solving tasks, still associated with uncertainties. This preliminary study should be understood as a proof-of-the-concept. The rules analysis has been revealed, and the cognitive process underlying this analysis has been discussed. This method allows a hierarchy in complex pathophysiology networks. Our brief overview of the current situation has shown the main direction of research activities in this domain – how to filter the generated set of rules effectively. One possibility is to apply suitable methods in the pre-processing phase like clustering, feature selection, domain knowledge provided by the expert, or stored in the knowledge base. The second possibility is to filter the resulting list by a suitable pruning approach or relevant metrics like confidence, support, or lift. In general, this direction should improve the initial results.

In this study, a greater emphasis in the development of mild cognitive impairment is put on more specific pathophysiology background, including immune system disorders and malnutrition, than on a broader clinical framework, based on chronic renal impairment and long-term hypertension. The extracted rules may help medical professionals recognize clinical patterns, although the method is approximatively, and the final decision depends on the expert. The CV risk factors are the best-known risk factors for MCI. The rules that contain only these factors (Tables 3 and 4) are therefore convincing to show how the Association rule mining method works in solving complex medical problems, such as aging-associated comorbidities. When the results of the analysis of the rules containing only CV risk factors

(Tables 3 and 4) and of those containing the whole dataset (Tables 5 and 6) were taken together, it was possible to recognize the two main clinical patterns, that are likely to be associated with MCI. The first one is specific for the younger patient group (63–73 years old), and the second one is specific for the older patient group (73 years old and more).

Finally, we want to stress the importance of cooperation between health services and computer science methods. This cooperation can be a win-win situation in which enough quality data will be available, common vocabulary will be defined, analytical results will be presented in a simple, understandable form, the knowledge will be applied into practice, and further integrated into a unique, reliable scenario.

New studies design is usually planned within the framework of the existing knowledge. Precisely this is an advantage of the exploratory studies such as this one, where the rule-based methods and large datasets are used to look for exciting new parameters and new concepts which are likely to go beyond the current theories. It contributes to the continuously improved and enhanced body of knowledge used in the medical diagnostics process. Specifically, in the case of mild cognitive impairment, early diagnostics is essential to prevent the serious decline of dementia. The obtained results helped us to design further research questions and experiments. In our future work, we will aim at providing a view on expected diagnostics that is as comprehensive as possible, considering all available aspects from various sides.

Acknowledgement

The research was partially supported by The *Slovak Research and Development Agency* under grants no. APVV-16-0213 and no. APVV-17-0550.

ORCID

František Babič  <http://orcid.org/0000-0003-2225-5955>

Ljiljana Trtica Majnarić  <http://orcid.org/0000-0003-1330-2254>

References

- Aevarsson, O., & Skoog, I. (2000). A longitudinal population study of the mini-mental state examination in the very old: relation to dementia and education. *Dementia and Geriatric Cognitive Disorders*, 11(3), 166–175. <https://doi.org/10.1159/000017231>
- Agrawal, R., & Srikant, R. (1994). Fast algorithms for mining association rules. In *Proceedings of the 20th International Conference on Very Large Data Bases*, (pp. 487–499). VLDB.
- Albert, M.S., DeKosky, S.T., Dickson, D.W., Dubois, B., Feldman, H.H., Fox, N.C., Gamst, A., Holtzman, D.M., Jagust, W.T., Petersen, R.C., et al. (2011). The diagnosis of mild cognitive impairment due to Alzheimer's disease: Recommendations from the National Institute on Aging-Alzheimer's Association workgroups on diagnostic guidelines for Alzheimer's disease. *Alzheimer's and Dementia*, 7(3), 270–279. <https://doi.org/10.1016/j.jalz.2011.03.008>
- Altman, N. S. (1992). An introduction to kernel and nearest-neighbor nonparametric regression. *The American Statistician*, 46(3), 175–185. <https://doi.org/10.2307/2685209>
- Alwidian, J., Hammo, B. H., & Obeid, N. (2018). WCBA: Weighted classification based on association rules algorithm for breast cancer disease. *Applied Soft Computing*, 62, 536–549. <https://doi.org/10.1016/j.asoc.2017.11.013>
- Antonopoulos, A.S., Margaritis, M., Lee, R., Chanook, K., & Antoniadis, C. (2012). Statins as anti-inflammatory agents in atherogenesis: Molecular mechanisms and lessons from the recent clinical trials. *Current Pharmaceutical Design*, 8(11), 1519–1530. <https://doi.org/10.2174/138161212799504803>
- Avila-Funes, J.A., Amieva, H., Barberger-Gateau, P., Le-Goff, M., Raoux, N., Ritchie, K., Carriere, I., Tavernier, B., Tzourio, C., Gutierrez-Robledo, L.M. et al. (2009). Cognitive impairment improves the

predictive validity of the phenotype of frailty for adverse health outcomes: the three-city study. *Journal of the American Geriatrics Society*, 57, 453–461. <https://doi.org/10.1111/j.1532-5415.2008.02136.x>


- Boban, M., Malojcic, B., Mimica, N., Vuković, S., Hof, P.R., & Simić, G.** (2012). The reliability and validity of the Mini-Mental State Examination in the elderly Croatian population. *Dementia and Geriatric Cognitive Disorders*, 33, 385–392. <https://doi.org/10.1159/000339596>
- Borah, A., & Nath, B.** (2018). Identifying risk factors for adverse diseases using dynamic rare association rule mining. *Expert Systems with Applications*, 33, 233–263. <https://doi.org/10.1016/j.eswa.2018.07.010>
- Buchanan, A. V.** (2006). Dissecting Complex Disease: The Quest for the Philosopher's Stone? *International Journal of Epidemiology*, 35(3), 562–571. <https://doi.org/10.1093/ije/dyl001>
- Bugnicourt, J. M., Godefroy, O., Chillon, J. M., Choukroun, G. & Massy, Z. A.** (2013). Cognitive disorders and dementia in CKD: The neglected kidney-brain axis. *Journal of the American Society of Nephrology*, 24 (3), 353–363. <https://doi.org/10.1681/ASN.2012050536>
- Cavagna, L., Boffini, N., Cagnotto, G., Inverardi, F., Grosso, V., & Caporali, R.** (2012). Atherosclerosis and Rheumatoid Arthritis: More Than a Simple Association. *Mediators of Inflammation*, 2012, ID 147354. <https://doi.org/10.1155/2012/147354>
- Deleidi, M., Jäggel, M., & Rubino, G.** (2015). Immune aging, dysmetabolism, and inflammation in neurological diseases. *Frontiers in Neuroscience*, 9, 172–178. <https://doi.org/10.3389/fnins.2015.00172>
- Futagami, S., Takahashi, H., Norose, Y., & Kobayashi, M.** (1998). Systemic and local immune responses against *Helicobacter pylori* urease in patients with chronic gastritis, distinct IgA and IgG productive sites. *Gut*, 43(2), 168–175. <https://doi.org/10.1136/gut.43.2.168>
- Harap, M., Husein, A. M., Aisyah, S., Lubis, F. R., & Wijaya, B. A.** (2018). Mining association rule based on the disease population for recommendation of medicine need. *Journal of Physics: Conference Series*, 1007(1). <https://doi.org/10.1088/1742-6596/1007/1/012017>
- Hogervorst, E., Huppert, F., Matthews, F. E., & Brayne, C.** (2010). Thyroid function and cognitive decline in the MRC Cognitive Function and Ageing Study. *Psychoneuroendocrinology*, 33(7), 1013–1022. <https://doi.org/10.1016/j.psyneuen.2008.05.008>
- Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C. R., & Wirth, R.** (2000). *CRISP-DM 1.0: Step-by-step data mining guide*.
- Jain, S., Gautam, V., & Naseem, S.** (2011). Acute-phase proteins: As diagnostic tool. *Journal of Pharmacy and BioAllied Sciences*, 3(1), 118–127. <https://doi.org/10.4103/0975-7406.76489>
- Kerber, R.** (1992). ChiMerge: Discretization of Numeric Attributes. In *AAAI'92 Proceedings of the Tenth National Conference on Artificial Intelligence* (pp. 123–128). ACM.
- Lakshmi, K. S., & Vadivu, G.** (2017). Extracting Association Rules from Medical Health Records Using Multi-Criteria Decision Analysis. *Procedia Computer Science*, 115, 290–295. <https://doi.org/10.1016/j.procs.2017.09.137>
- Lazarczyk, M. J., Hof, P. R., Bouras, C., & Giannakopoulos, P.** (2012). Preclinical Alzheimer Disease: Identification of Cases at Risk among Cognitively Intact Older Individuals. *BMC Medicine*, 10(1), 127–135. <https://doi.org/10.1186/1741-7015-10-127>
- Lloyd, S. P.** (1982). Least squares quantization in PCM. *IEEE Transactions on Information Theory*, 28(2), 129–137. <https://doi.org/10.1109/TIT.1982.1056489>
- Qiu, S., Chang, G. H., Panagia, M., Gopal, D. M., Au, R., & Kolachalama, V. B.** (2018). Fusion of deep learning models of MRI scans, Mini-Mental State Examination, and logical memory test enhances diagnosis of mild cognitive impairment. *Alzheimer's & Dementia: Diagnosis, Assessment & Disease Monitoring*, 10(1), 737–749. <https://doi.org/10.1016/j.dadm.2018.08.013>
- Postiglione, A., Milan, G., Ruocco, A., Gallotta, G., Guiotto, G., & Di Minno, G.** (2001). Plasma Folate, Vitamin B (12), and Total Homocysteine and Homozygosity for the C677T Mutation of the 5,10-Methylene Tetrahydrofolate Reductase Gene in Patients with Alzheimer's Dementia. A Case-Control Study. *Gerontology*, 47(6), 324–329. <https://doi.org/10.1159/000052822>
- Roberts, R. O., Geda, Y. E., Knopman, D. S., Boeve, B. F., Christianson, T. J., Pankratz, V. S., Kullo, I. J., Tangalos, E. G., & Petersen, R. C.** (2009). Association of C-Reactive Protein with Mild Cognitive Impairment. *Alzheimer's & Dementia*, 5(5), 398–405. <https://doi.org/10.1016/j.jalz.2009.01.025>
- Sariyer, G., & Öcal Taşar, C.** (2020). Highlighting the rules between diagnosis types and laboratory diagnostic tests for patients of an emergency department: Use of association rule mining. *Health Informatics Journal*, 26(2), 1177–1193. <https://doi.org/10.1177/1460458219871135>

- Shearer, C.** (2000). Strategic Modeling for the Characterization of the Conditions That Allow the Anticipation of the Consumer's Requests. The CRISP-DM Model: The New Blueprint for Data Mining. *Journal of Data Warehousing*, 5, 13–22. <https://doi.org/10.4236/jss.2015.310021>
- Schrijvers, E. M. C., Witteman, J. C. M., Sijbrands, E. J. G., Hofman, A., Koudstaal, P. J., & Breteler, M. M. B.** (2010). Insulin Metabolism and the Risk of Alzheimer Disease: The Rotterdam Study. *Neurology*, 75(22), 1982–1987. <https://doi.org/10.1212/WNL.0b013e3181ffe4f6>
- Tjia, J., Velten, S.J., Parsons, C., Valluri, S., & Briesacher, B.A.** (2010). Studies to reduce unnecessary medication use in frail older adults: a systematic review. *Drugs Aging*, 30, 285–307. <https://doi.org/10.1007/s40266-013-0064-1>
- Umegaki, H.** (2014). Type 2 Diabetes as a Risk Factor for Cognitive Impairment: Current Insights. *Clinical Interventions in Aging*, 9, 1011–1019. <https://doi.org/10.2147/CIA.S48926>
- Van Guldener, C.** (2006). Why Is Homocysteine Elevated in Renal Failure and What Can Be Expected from Homocysteine-Lowering? *Nephrology Dialysis Transplantation*, 21(5), 161–1166. <https://doi.org/10.1093/ndt/gfl044>
- Walker, S. R., Wagner, M., & Tangri, N.** (2014). Chronic kidney disease, frailty and successful aging: a review. *Journal of Renal Nutrition*, 24(6), 364–370. <https://doi.org/10.1053/j.jrn.2014.09.001>
- Whaley-Connell, A., & Sowers, J.R.** (2018). Insulin resistance in kidney disease: Is there a distinct role separate from that of diabetes or obesity. *Cardiorenal Medicine*, 8(1), 41–49. <https://doi.org/10.1159/000479801>
- Whitmer, R. A.** (2007). The Epidemiology of Adiposity and Dementia. *Current Alzheimer Research*, 4(2), 117–122. <https://doi.org/10.2174/156720507780362065>
- Zhang, X., Hu B., Ma X., Moore P. & Chen, J.** (2014). Ontology Driven Decision Support for the Diagnosis of Mild Cognitive Impairment. *Computer Methods and Programs in Biomedicine*, 113(3), 781–791. <https://doi.org/10.1016/j.cmpb.2013.12.023>



Copyright © 2020 by the author(s). Licensee Prague University of Economics and Business, Czech Republic. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution License (CC BY), which permits use, distribution and reproduction in any medium, provided the original publication is properly cited, see <http://creativecommons.org/licenses/by/4.0/>. No use, distribution or reproduction is permitted which does not comply with these terms.

The article has been reviewed.

Editorial record: First submission received on 29 June 2020. Revisions received on 5 August 2020 and 19 August 2020. Accepted for publication on 20 August 2020. The editor in charge coordinating the review of this manuscript and approving it for publication was Zdenek Smutny .