Satakunnan ammattikorkeakoulu
Satakunta University of Applied Sciences

JUSSI BERGMAN

Data transformation, data conversion and custom tool development for Pori-75 study

DEGREE PROGRAMME IN ARTIFICIAL INTELLIGENCE
2021

Author(s) Bergman, Jussi	Bachelor's thesis	31.9.2021
	Number of pages	Language of publication: English
Title of publication		
Degree Programme BA in Artificial Intelligence		

Abstract

Objective : To describe the automated dimension engineering process developed specifically for Pori-75 data by using different data conversion methods and measurement scale transformations. The first objective was to prepare the Pori-75 data for effective use of Data-analytic(DA) and Artificial Intelligence(AI) tools. The second objective was to describe the developed transformation tools for blood result categorisation, and for ATC-coding of medication lists. Third objective was to develop a automatic data selection tool and a fast hypotheses testing tool specifically for Pori-75 data.

Methods : Data conversions and measurement scale transformations is utilised to the baseline data of 313 medical dimensions from 518 participants in Pori-75 study in year 2020. A tool for "quick-and-dirty" idea/hypotheses testing is developed by using automatic group mean comparison method together with several automatic statistical test and statistics tools. All processes is programmed with Python by using industry standard, open source and state-of-art libraries including SciPy, NumPy, Pandas, MatPlotLlib, SKLearn, Pingouin and Keras. This allows the code to be integrated to custom environments in the future.

Results : Original 313 data dimensions was augmented to over 1000 dimensions ready to be used with the latest DA and AI tools. The data conversion and transformation processes was programmed so that new data can be easily integrated yearly in automated basis. Python functions was developed for quick data selection and a fast idea/hypotheses testing. Python functions was created for automatic blood result categorisation and medication to ATC-code conversion.

Key words
data processing, statistical methods, Python (programming languages), medical data

CONTENTS

LIST OF SYMBOLS AND TERMS

Dimension :

In this study the term **dimension (data dimension)** is used interchangeably with **data feature** and **data attribute**. Intuitive alternative would be a column in a excel sheet.

Predictor :

In this study the term **predictor** or **predictor variable** means any **dimension** that is used for studying it's effect to a **target**

Target :

In this study term **target** means any **dimension** that is used for studying **predictors** effect to it. Term is interchangeable with **label** in supervised learning settings.

Measurement types :

In this study the term **measurement type** defines the **scale** of a dimension. Measurement types used in this study are: **nominal**, **ordinal**, **interval** and **ratio**. Also extra type is used : **dichotomous**

Categorial :

In this study the term **categorial** (**categorial dimension**) is used interchangeably with term **discrete (discrete dimension).** It represents characteristics that can be counted but not measured. Includes **nominal** and **ordinal** measurement typles.

Continuous :

In this study the term **continuous (continuous dimension)** represents measurements that can be measured but not counted. Includes **interval** and **ratio** measurement types.

Category :

In this study the term **category (data category)** represents the differen values found in a **categorial dimension**

1 INTRODUCTION

In June 2019, Social Security Center of Pori (Petu) launched the Pori75 counseling service, where everyone living in their homes in Pori co-operation area and who have also reached or will reach the age of 75 this year is called for a health examination. In 2016, almost 90% of the 75 year olds lived at their own homes. The goal is to increase the number of people living at home to 94% by 2025. The Pori-75 study will identify methods for detecting the health risks of older people in time for further research and treatment (Holm et al., 2021).

The Pori-75 research team is a multidisciplinary research group bringing together healthcare professionals with data processing, data analysis, statistics and AI - professionals. The safe technical environment with high end data lake provider and the possibility for high performance computing makes Pori 75 project globally unique.

Efficient use of multi-dimensional data in a multidisciplinary team requires careful and thorough data cleaning, conversion and transformation processes. Also, data-specific tools for easy and fast data selection and idea/hypothesis testing needs to be developed.

My role in the team is to act as a lead data-analyst, statistician and AI-professional. I also programmed all processes mentioned here with Python together with several open-source, industry-standard and state-of-art libraries. My role also includes supervision of a junior data scientist in cleaning the data and providing basic statistics. The data collection was done by other data-professionals in the team.

In this study I describe the dimension engineering process that I made for Pori-75 data by using different data conversion methods and measurement scale transformations. I also developed tools for blood test result categorisation, ATC-coding of medicine lists, a data selection tool and a fast hypotheses testing tool which will be covered with use case examples.

The said work is a mandatory step for using the Pori-75 data efficiently with modern machine learning tools. Modern AI and data-analytic tools are required for developing predictive models, meters, recommendation tools, visualisation tools, and many other ambitious products and tools in the scope of this research project.

No results relating to Pori-75 has yet been published. The data and the results remain confidential and only selected screen shots from the early results will be shown in this document with the permission of the the head of the AI Research group.

2 DATA DESCRIPTION

The Pori-75 data used in this study includes health examinations of 518 participants from year 2020. The data is comprised of 313 original data dimensions(Holm et al., 2021)
The data can be devided into 6 segments; Test surveys and meters, Measurements, Other patient information, Referrals, Blood test results and Medication lists.

Test surveys and meters includes 15 mostly Likert-scale questionaries (Joshi et al., 2015) covering altogether 153 health related questions. **Measurements** includes 8 basic health measurements including, weight, height, vision, hearing waist and neck circumfences. **Other patient information** is a set of 23 patient details covering many aspects from housing type to last dental visit. **Referrals** contains the information if a customer has been re-directed to another specialist during Pori-75 inspections. **Blood test results** cover 15 blood tests for each participant. **Medication lists** includes the ATC-coded medications (*Guidelines for ATC Classification and DDD Assignment 2021*, n.d.) for each participant.

The full overview of the original data is presented in Table 1.

The basic cleaning of the data was done by junior data-scientist under my supervision. Outlier removal was done after consultations with healthcare professionals. Inputing missing values was not required by the tools used in this study. Missing value imputation will be done with the optimal strategy dictated by the DA/AI -tool used in each scenario.

Table 1. The overview of the original data used in the study

Data	dimensions	Description
Test, survey or meter : 15D(15), ADL(7), IADL(8), UDI-6, GDS-15, SARC-F(5), AUDIT-C(3), LOTTA(20), Ortostatic Test & TOUNOU(chair test)(12), FRAIL(6), MNA(20), MMSE(20), STOP-BANG(11), MALLAMPATI(1) FROP-Com. (4)	153	Participants went throuhg 15 questionaries and tests: Health-related quality of life survey (15D) Functional capability (ADL and IADL) Basic Nordic Sleep Questionnaire (BNSQ) The Geriatric Depression Scale (GDS-15) Urogenital Distress Inventory (UDI-6) Alcohol Use Disorders Identification Test (AUDIT-C) A simple fraility questionare (FRAIL-Scale) Mini Nutritional Assesment (MNA) Mini-Mental State Examination (MMSE) STOP-BANG (**Snoring, Tiredness**, Observed apnea, blood Pressure, Body mass index, Age, Neck circumference and Gender) Falls Risk for Older People in Community setting (FROP-Com) Ortostatic test (ORTOS) Chair-Stand Test as a Measure of Lower Body Strength (TOUNOU) Mallampati score (MALLAMPATI) Medication security questionare (LOTTA)
Measurements : Vision (E-board), Hearing (Whisper test), Waist circumference, Neck circumference, Weight, Height,	8	Values necessary for disease identification and research was examined and measured from participants
Other patient info: Postal codes 3_areas, Postal codes: 5_areas, Postal_code, Gender, Marrietal_status, Education, Form_of_housing, house_type, House_with_elevator, House_with_stairs, Allergies, Mobility_aids, Special_diet, smoking, outside_help_required, drug_dosing_by, drug_dosing_method, entitled_for_care_allowance(hoitotuki), Drivers_licence, Teeth_brushing_freq, Teeth_denture, Teeth_last_dental_visit	23	Extensive patient info was collected.
Referrals : Doctor, Dentist, Physiotherapist, Pharmacist, Nurse, Nutritionist, Memory Manager, Diabetes nurse, Psych nurse, Service instructor, Respiration nurse, Geriatric policlinic, A-clinic, Hearing test.	14	Record if participant was referred to another medical-appointment from Pori-75 reception visit.
Blood tests : B-Leuk, B-Hb, B-Hkr, B-Eryt, E-MCV, E-RDW,E-MCH, E-MCHC, B-Trom, HbA1c, D-25-OH, Krea (GFR), Na, K, Ca-albK	15	Participants went through blood tests including; total number of leukocytes, hemoglobin, hematocrit, red blood cell volume, number of erythrocytes, mean erythrocyte volume, variation in red blood cell size, mean red cell hemoglobin, mean red blood cell hemoglobin concentration, number of platelets, Sugar Hemoglobin, Vitamin D, Creatinine, Sodium, Potassium and Albumin Corrected Calcium
Medications-list	108	Patients medications were mapped and medication lists were updated.
sum	**313**	

3 DATA CONVERSIONS

3.1 Replacing numeric nominal, ordinal and dichotomous values with textual category names/descriptions

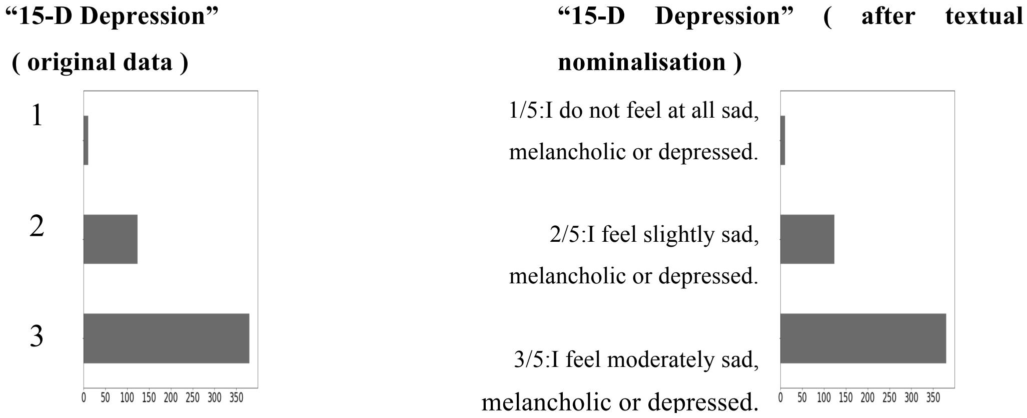
Data received from data collection process is often purely in numerical form. This was the case with Pori-75 data as well. Numeric form is the natural form for continuous dimensions, that is, interval and ratio measurement scales but for categorial variables numbers dont tell much.

For example questionaries have questions usually in written form like in the 5-point Likert-scale questions in 15D quality of life meter (Sintonen, 1995). The example question below is about measuring depression :

1 ( ) I do not feel at all sad, melancholic or depressed.
2 ( ) I feel slightly sad, melancholic or depressed.
3 ( ) I feel moderately sad, melancholic or depressed.
4 ( ) I feel very sad, melancholic or depressed.
5 ( ) I feel extremely sad, melancholic or depressed.

This type of answer is usually recorded as number 1-5. By doing this the recorder mistakingly removes the information needed for easy clinical interpretation as can be seen from image 1. This is why the original questionaries for each of the 15 questionaries used in Pori 75 was collected and values in the corresponding dimensions was replaced with a text describing the original textual answer. Altogether 153 questionary questions was converted and also translated in english.

Image 1. Example of visualisation before and after replacing nominal and ordinal values with descriptive text.



“15-D Depression” (original data)

“15-D Depression” (after textual nominalisation)

3.2 The dimension name and category naming conventions

The main purpose of informative and consistent dimension and category naming convention is to carry all required information about the original data, question and the measurement scale through the whole process from data conversion to the results and visualisations provided to healthcare professionals.

In Pori-75 data, dimension name and/or category should include :

- the name of the questionary/measurement (“Lotta” / “B-Trom(2791)”),
- the number of the question in questionary (_2_)
- Text that discloses the answer in understandable way (“gets treatment from more than one doctor”)
- Text-string disclosing information about the scale used (“1Yes_2No_dich)

The naming conventions used in Pori-75 data conversion process is presented in Table 2

Table 2. The naming convention examples used in Pori-75 data conversion process

	Categorial dimensions	Continuous dimensions
Original questionary / inspection name	"Lotta" : question number 2	blood test : B-Trom(2791)
Original question/result	Question 2 : "Have you been treated by more than one doctor during the last year?"	No question, simply blood test result
String	**"Lotta_2_more_than_one_doctor"**	**"B-Trom(2791)"**
category names	YES = "gets treatment from more than one doctor" NO = "Does not get treatment from more than one doctor"	Ratio transformed (categorised) to nominals based on clinical tresholds : at_least_one_value abowe_limits_in_insp_year at_least_one_value in_limits_in_insp_year at_least_one_value below_limits_in_insp_year at_least_one_value abnormal_in_insp_year
Category measurement type	Dichotomous (Dich)	Dichotomous (Dich)
String 2	"_1Yes_2No_dich"	"_1Yes_2No_dich"
Final category name	" **gets_treatment_from_more_than_one_doctor_1Yes_2No_Dich** "	B-Trom(2791)_at_least_one_value abowe_limits_in_insp_year_1Yes_2No_dich **OR** B-Trom(2791)_at_least_one_value in_limits_in_insp_year_1Yes_2No_dich **OR** B-Trom(2791)_at_least_one_value below_limits_in_insp_year_1Yes_2No_dich **OR** B-Trom(2791)_at_least_one_value abnormal_in_insp_year_1Yes_2No_dich

3.3 Medications list categorisation by using ATC-coding for nominal-ordinal conversion

WHO maintains a Anatomical Therapeutic Chemical (ATC) code for medicine classification where active substances are divided into different groups according to the organ or system on which they act and their therapeutic, pharmacological and chemical properties(*Guidelines for ATC Classification and DDD Assignment 2021*, n.d.).

Pori 75-study includes medication lists collected from participants. Medication lists needed to be categorised based on ATC-coding.

A Python function was developed for converting medication lists to ATC-codes automatically with optionality for defining the number/depth of ATC levels to use.

4 MEASUREMENT SCALE ANALYSIS

Statistics, data-analytics and AI offers plethora of mathematical tools for retrieving information about the data and tools to build models for classification and regression. To be able to use such tools, we need to first investigate the data thoroughly and to understand the origin of each dimension.

The most important step in understanding the data is to analyse the data measurement scale for each dimension. The measurement scale type together with the distribution of the dimension dictates which mathematical, statistic and machine learning tools can be used for each dimension(Mishra et al., 2019).

All data can be measured in nominal, ordinal, interval or ratio scales. This means the dimension can have information about the category, order, distance and/or zero origin respectively(Joshi et al., 2015).

In data the measurement scale of each dimension is dictated by the source of the information. A source of the information might be a nurse asking a patient about allergies(nominal), filling a Likert-scale questionary about nutrition(ordinal) or a doctor measuring the body temperature in celsius(interval) or weight(ratio) .

For example a dimension including list of allergies for each patient is in nominal scale with information only about the category of the allergy. It has no information about the order between the categories, that is, birch allergy is no better or worse than lactose intolerance or any food allergy. These are simply nominal categories without information about the order between them. There is no rational in using any other mathematical tools with nominal values than mode and frequency.

On the other hand, dimension including the answers from a Likert-scale questionary includes information about the order as well. Likert scale questionaries allow user to choose the answer from a ordinal scale : eg. “Strongly Agree “- “Agree” - “Undecided” - “Disagree” - “Strongly Disagree”, in where “Agree” is somewhat more than “Disagree” and “Undecided” lays in between these two.

Ordinal information from Likert-scale questionary has information about the category and the order but does not have equal distances between the categories. Body temperature which is measured in celsius has equal distance between each degree. This is why temperature(celsius) is in interval scale. There is no sense in summing or multiplying the values in ordinal scale but with interval scale like temperature(celsius) it makes sense. Once again the scale defines which mathematical tools should be used. With interval scale, summing the values might make sense but multiplication does not because there is no true zero (-3℃ x +3℃=?). Zero is a temperature as much as -13 or +25.

Unlike interval scale, the ratio scale has also the “true zero”. Weight and height are good examples of ratio scale where zero weight truly means no weight at all. Interval scale allows using mathematical tools quite freely(Mishra et al., 2019).

In addition to the mentioned basic measurement scales a separate Dichotomous data type is often used. It doesn’t take a stand about measurement scale as such. It is only used when a dimension has precisely two categories. A dimension having only two categories is nominal or ordinal by its nature but has many qualities similar to ratio values when the category names are converted as 0 and 1 and a mean is calculated. In this form a dichotomous dimension type behaves as a frequency probability for the two categories and inherits many of the characteristics from ratio measurement type. This allows many tools to be used with dichotomous variables with varying limitations.

The reasoning between measurement scales and the DA and AI - tools used in this study is presented in table 2.

Table 2. The tools used in the study categorized by the dimension measurement type

	nominal	**ordinal**	**interval**	**ratio**	**dichotomous**
EXAMPLES	1=Single, 2=Married, 3=Divorced, 4=Widowed	1=Disagre 2=Neither agree nor disagree, 3=Agree,	pulse, weight, length,	temperature(celsius), scores_between_0-1	sex, 1=yes, 2=no, any dimension with only 2 categories
SCALE MEASUREMENT PROPERTIES					
Class / Identity	X	X	X	X	X
Order Magnitude	-	X	X	X	With limitations
Equal intervals	-	-	X	X	With limitations
Minimum value is zero	-	-	-	X	With limitations
MATHEMATICAL TOOLS					
Deduction (-) and addition (+)	-	-	X	X	X
Multiplication (x) and division (/)	-	-	-	X	**-**
Mode	X	X	X	X	**X**
Median	-	X	X	X	**-**
Mean	-	-	X	X	With limitations
Pearsons correlation	-	-	X	X	With limitations
Spearman's correlation	-	X	-	-	With limitations
chi-square-test	-	X	-	-	X
ANOVA/t-test	-	-	X	X	With limitations
distance based similarity	-	-	X	X	**-**
DA / AI-TOOLS					
Classification (predictor/independent)	X	X	With limitations	With limitations	X
Classification (target/dependent)	X	X	With limitations	With limitations	X
Regression (predictor/independent)	With limitations	With limitations	X	X	With limitations
Regression (target/dependent)	-	-	X	X	With limitations
Clustering	-	-	X	X	**-**
mean based group analysis (predictor/independent	X	-	-	-	X
mean based group analysis (target/dependent)	-	-	X	X	With limitations

5 MEASUREMENT SCALE TRANSFORMATIONS.

Many Machine learning tools benefit from using categorial dimensions as a target, predictor or both. Some other tools perform better with continuous input an/or output. Also, a mean based group analysis that is used in this study(A. Breck AND B. Wakar, 2021)(Sleight, 2000) is very powerful and is dependent purely on nominal dimensions.

Many medical measures have the problem of “variability within the individual and within the laboratory.” E.g blood test values vary by age, sex and within the menstrual, diurnal and seasonal cycles(Phillips, 2009). This is why many continuous medical measures requires categorisation to make the results comparable. Adding categorised medical dimensions alongside the original continuous dimensions mitigates these problems.

For these reasons it is beneficial to add new dimensions to the data having all reasonable transformations of the original measurement scale. Even though some methods exists for quite reliable transformation of categorial data to continuous(Marateb et al., 2014), in this study no new information was added to the original data without clinical consultation and validated references. This is the case e.g. when counting the MNA-score in pre-defined manner(Nestle Nutrition Institute, n.d.) or when weighting the answers in 15D(Sintonen, 1995) to make the ordinal dimension behave like in ratio scale.

Backward transformations eg. from ratio to ordinal or ordinal to nominal doesn’t introduce new information but rather information is lost. This is why the transformed dimensions are added alongside the original dimensions and thus no information is lost in the data.

All transformations and conversions used in this study is described in Table 3**.**

Table 3. Data transformation and conversion methods used in the study

original type	converted type	primary conversion/ transformation method	Example	secondary method	Example
Interval / Ratio	ordinal	clinical tresholds : normal ranges	Blood tests : B-HbA1c: <20=below normal, 20-42=normal, >42=above normal	Binning / percentiles	0-35%=Below average, 33-66%=average, 66-100%=above average
nominal	nominal	Pre-defined categories (e.g ATC-coding)	Medication: ATC-coding : losartan=C09, marevan=B01 …		
Interval / Ratio / ordinal	nominal	replacing values with descriptive text	15D-"Depression" : 1= I do not feel at all sad, melancholic or depressed. 2=I feel slightly sad, melancholic or depressed. 3=:I feel moderately sad, melancholic or depressed ...		
ordinal / nominal	ratio	Calculating scores based on clinical references	calculating 15D-score based on weights defined in medical publications.		

5.1 Blood results categorisation and medical data - using clinical thresholds for ratio to ordinal transformation

Pori 75 data has 15 dimensions of blood test results. These dimensions was converted to ordinal dimensions by binning the values to “below normal”, “normal”, “above normal” and “abnormal”.

The thresholds for each dimension, taking account all necessary features like sex and age, was determined by healthcare professionals.

The “below normal”, “normal” and “above normal” creates a ordinal scale.

A nominal category “abnormal” was added after exploratory analysis showing significant prediction power.

5.2 Medical data - using binning/percentiles for ratio-ordinal transformation

If no clear clinical threshold was available for a dimension like e.g for chair_test_pulse_recovery and for UDI-score, a secondary strategy of simple binning to four (25%) percentiles was used.

5.3 Scoring

Questionaries usually have a pre-defined way of calculating the scores. If no reference is available, a simple additive scoring was used.

All original and converted new dimensions of Pori-75 data is presented in Table 4.

Table 4. Original and converted new dimensions of Pori-75 data (2020), n=518.

Data overview	**Orig. Dim.**	**Nominal Categories as descriptive text**	**Numeric Ordinals**	**Continu ous**	**Numeric Dichotom ous**
Test, survey or meter : 15D, ADL(7), IADL(8), UDI-6, GDS-15, SARC-F(5), AUDIT-C(3), LOTTA(20), Ortostatic Test & TOUNOU(chair test)(12), FRAIL(6), MNA(20), MMSE(20), STOP-BANG(11),MALLAMPATI(1) FROP-Com. (4)	153	255	90	17	46
Measurements : Vision (E-board), Hearing (Whisper test), Waist circumference, Neck circumference, Weight, Height,	8	8	1	5	2
Other patient info: Postal codes 3_areas, Postal codes: 5_areas, Postal_code, Gender, Marrietal_status, Education, Form_of_housing, house_type, House_with_elevator, House_with_stairs, Allergies, Mobility_aids, Special_diet, smoking, outside_help_required, drug_dosing_by, drug_dosing_method, entitled_for_care_allowance(hoitotuki), Drivers_licence, Teeth_brushing_freq, Teeth_denture, Teeth_last_dental_visit	23	248	2	0	36
Referrals : Doctor, Dentist, Physiotherapist, Pharmacist, Nurse, Nutritionist, Memory Manager, Diabetes nurse, Psych nurse, Service instructor, Respiration nurse, Geriatric policlinic, A-clinic, Hearing test.	14	14	0	0	14
Blood tests : PVK, HbA1c, D-25-OH, Krea (GFR), Na, K, Ca-albK (cat: below limits, in limits, above limits, abnormal_y)	7	28	4	0	0
Medications-list (ATC :A-V & A01-V20)	108	108	0	0	108
sum	**313**	**661**	**97**	**22**	**198**
				Tot.Dim	**978**

6 PORI-75 DATA-SPECIFIC TOOL FOR FAST DATA SELECTION AND HYPOTHESIS TESTING BY COMPARING INDEPENDENT GROUP MEANS.

6.1 Data selection tool

After data conversion and transformation processes the Pori-75 data has over 1000 dimensions ready to be used with different DA and AI tools.
The Pori-75 project also has several data-related research ideas and hypotheses and new ideas born on weekly basis. Each scenario requires different AI/DA tools with different measurement scale requirements for both predictors and targets. This creates a need for a way to quickly select the dimensions with optimal measurement scale transformations suitable for each DA/AI tool.

For example regression and clustering tools works optimally with interval or ratio data, classification tools with nominal or ordinal data, group mean based tools with nominal data and neural network models have variety of measurement scale requirements based on the model.

Statistical methods, bayesian analysis, optimisation tools and many other custom mathematical tools have their own requirements for measurement scale and/or distribution. Overview of these requirements can be found from Table 2.

For these reasons Python functions for quickly selecting predictor dimensions and target dimensions was developed. The selection table presented here is a high level selection tool customised for group mean comparison but can easily be customised to other tools as well.

A visualisation of the data selection table is presented in Table 5.

Table 5. The selection table for selecting predictors, targets, settings and mathemetical tools for quick hypotheses testing with Pori75 transformed data.

Predictors		Targets		Settings		
15D		15D		Create .xlxs		X
ADL		ADL		count scores for all tests		
IADL		IADL				
UDI-6		UDI-6		**TOOLS**	**ABBR**	
GDS-15		GDS-15		difference between predictor and target	**abs-diff**	X
SARC-F		SARC-F		difference between predictor and target (percentage)	**%-diff**	X
AUDIT-C		AUDIT-C		absolute difference between predictor and target (percentage)	**abs_%-diff**	X
ORTOS		ORTOS		p-value of the T-test for the means of two independent samples of scores	**ttest_pval**	X
TOUNOU		TOUNOU		statistics of the T-test for the means of two independent samples of scores	**ttest_stat**	X
FRAIL		FRAIL		Bayesian likelihood ratio	**bayesian_F**	X
MNA		MNA		target mean for the group		X
MMSE		MMSE		target mean for the whole sample_(selected subgroup included)		X
STOP-BANG		STOP-BANG		target mean for the whole sample (subgroup not included)		X
MALLAMPATI		MALLAMPATI		confidence intervals	**ci95_lo,** **ci95_hi**	X
FROP-Com		FROP-Com		minimum value in a group	**min**	X
Blood Tests		Blood Tests	X	maximum value in a group	**max**	X
LOTTA	X	LOTTA		group variance	**var**	X
Medications (ATC)		Medications (ATC)	X	group mean absolute deviation	**mad**	X
Referrals		Referrals		group standard deviation	**std**	X
Measurements		Measurements		group standard error	**ste**	X
Patient Info		Patient Info				
Postal codes		Postal codes	X			

6.2 Quick-and-dirty hypothesis testing by comparing independent group means

Typical hypothesis in projects like Pori-75 relates to questions if a certain predefined dimension predicts another one or if they are in relationship to each other. One example is to explore if medication risk measured with LOTTA-questionary relates to the ATC-codes in the actual medication lists or to blood test results. And maybe if there are differences between different postal code areas. Sometimes the goal is to use DA/AI tools to find the optimal dimensions that can be used to predict another dimension e.g risk in falling, probability for different diagnoses and risk of being hospitalised, just to mention few.

With high-dimensional data with varying distributions and measurement scales, Pearson correlation can't always be used to reliably describe such relationships(Kim, 2018). Also, the fundamentals underlying Pearson correlation and significance testing build upon variance(Gorard, 2005) creates difficulties for some scientists in trusting such methods and results(Dienes & Mclatchie, 2017). Said that, traditional statistics used in science is based on variance.

Selecting the right methods for reliable hypothesis testing depends also on the distribution, measurement scale, size of the group and dimensional independency(Upton & Cook, 2008). Some measurement scale specific compromises can be made safely like using t-test with Likert-scale data(Winter & Dodou, 2019), but the best way to satisfy the needs of all scientists with differing theoretical backgrounds is to provide statistics from several test-methods simultaneously.

For these reasons a group mean analysis(A. Breck And B. Wakar, 2021) was selected as a first tool for inspecting relationships. The possibility to select from variety of statistics and hypothesis testing methods including t-tests, t-statistics, confidence intervals, Mann-Whitney-Wilcoxon and bayesian factor was provided. Also, all group means required for estimating the relationships between group means is provided.

6.2.1 group means comparison method

Comparing group means method used in this study can be understood through the following steps. An example is provided in Figute 3

STEP 1 : Select the dimension to use as predictor (BMI)

STEP 2 : Select the dimension to use as target (Quality of life score)

STEP 3 : calculate the target **mean** (quality of life mean) **for all data** (ALL=0,721)

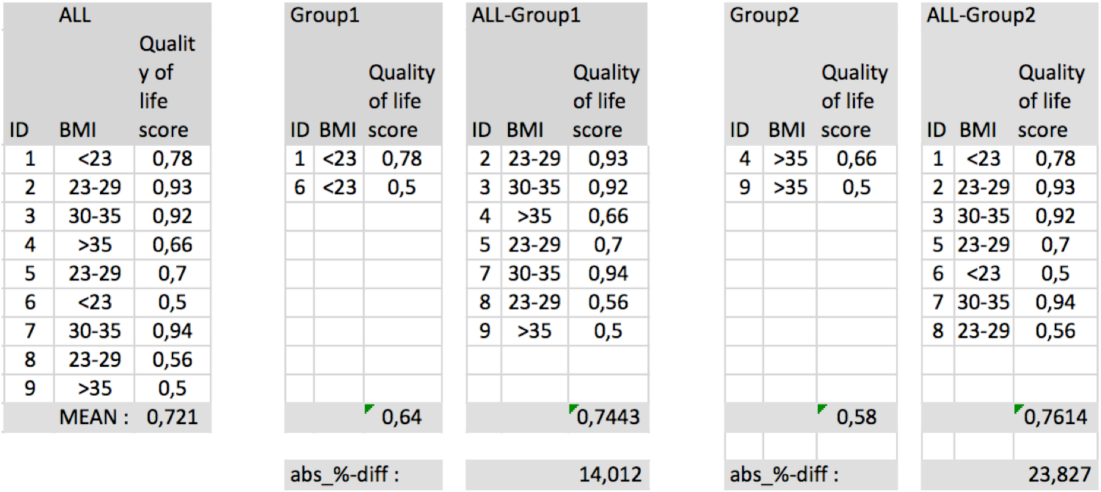
STEP 4 : Select first group for comparison (e.g. all participants with BMI below 23, Group1=BMI<23)

STEP 5 : calculate the target **mean for the group** (Group1_mean=0,64)

STEP 6 : calculate the target **mean for All data without participants in Group1** (ALL-Group1)

STEP 7 : compare the means

STEP 8 : Calculate statistics and tests



ALL

ID	BMI	Quality of life score
1	<23	0,78
2	23-29	0,93
3	30-35	0,92
4	>35	0,66
5	23-29	0,7
6	<23	0,5
7	30-35	0,94
8	23-29	0,56
9	>35	0,5
	MEAN :	0,721

Group1

ID	BMI	Quality of life score
1	<23	0,78
6	<23	0,5
		0,64
abs_%-diff :		

ALL-Group1

ID	BMI	Quality of life score
2	23-29	0,93
3	30-35	0,92
4	>35	0,66
5	23-29	0,7
7	30-35	0,94
8	23-29	0,56
9	>35	0,5
		0,7443
		14,012

Group2

ID	BMI	Quality of life score
4	>35	0,66
9	>35	0,5
		0,58
abs_%-diff :		

ALL-Group2

ID	BMI	Quality of life score
1	<23	0,78
2	23-29	0,93
3	30-35	0,92
5	23-29	0,7
6	<23	0,5
7	30-35	0,94
8	23-29	0,56
		0,7614
		23,827

Figure 3. The process of comparing group means

If the difference between the means is tested as significant, it can be considered as an indication of relationship between the dimensions and/or groups. This indication needs to be verified by using e.g. multivariable regression before any reporting to clinicians to avoid misleading results typical for such method.

The algorithm developed for this study uses simple “brute-force” for iterating through all combinations of 2 dimensions specified in the data selection table by comparing the means and calculating all tests and statistics defined by the user.

6.2.2 assessing results found through group means comparison

In this study, comparing group means is used only as a “quick-and-dirty” tool for identifying potential relationships based on pre-hoc hypothesis and all result will be confirmed with e.g. multivariable regression.

Comparing group means, also called subgroup analysis, is an efficient method for finding early indications of relationships between dimensions and categories. Without careful statistical evaluation, comparing group means easily produces misleading information by proposing relationships that are not really there.
This information easily confuses clinicians and directs the research project to directions that are not based on pre-hoc hypotheses any more(Sleight, 2000)(A. BRECK AND B. WAKAR, 2021). A credible subgroup analysis needs to meet specified credibility criteria listed below(Inglis et al., 2018).

- Is the subgroup variable a characteristic measured at baseline?
- Was the subgroup variable a stratification factor at randomisation?
- Was the hypothesis specified a priori?
- Was the subgroup analysis one of a small number of subgroup analyses tested (≤5)?
- Was the test of interaction significant (interaction $p < 0.05$)?
- Was the significant interaction effect independent, if there were multiple significant interactions?
- Was the direction of the subgroup effect correctly pre-specified?
- Was the subgroup effect consistent with evidence from previous studies?
- Was the subgroup effect consistent across related outcomes?
- Was there indirect evidence to support the apparent subgroups effect (biological rationale, laboratory tests, animal studies)?

All results in Pori-75 is verified by using e.g. multivariable regression before any reporting to clinicians. Also, the p-value requirements reported in Inglis et al. should not be fixed ($p < 0.05$) but rather be proportional to the size of the full data. In Pori-75-data with ≈1000 subgroups the p-value needs to be; $p< 0.001$. Also, in Pori-75 study the p-value is always assessed together with bayesian factor and subgroup size which needs to be >10.

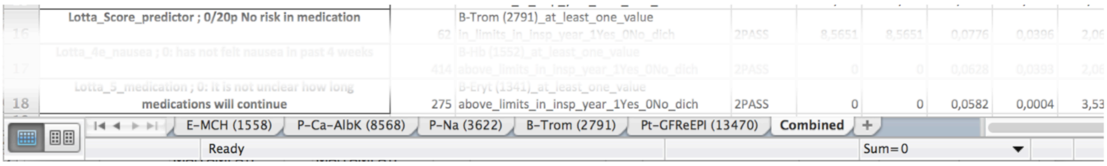
6.3 Example use case : inspecting potential relationship between medication symptoms questionary(LOTTA) and blood test results, medication lists and postal codes.

Example of a hypothesis testing is to inspect if medication risk measured in LOTTA-questionary relates to the ATC-codes in the actual medication lists or to blood results. And maybe there are differences between different postal code areas. The steps in using the Pori-75 tool are the following :

Step 1 : select LOTTA as the predictor form the selection table.
Step 2 : select Blood Tests, Medications(ATC) and Postal Codes as Targets.
Step 3 : Because we have more than one target, select “create.xls” from the settings in selection table
Step 4 : select the tools for statistics and testing from selection table
Step 5 : run the process
Step 6 : open the resulted .xlxs file to Excel and find the statistics for all targets in their own tabs (Figure 3).
Step 7 : Manually evaluate results from Data-Analytic and statistics perspectives (Figure 4).
Step 8 : If clear evidence of potential relationship is found, confirm the results.
Step 9 : Present the results for clinicians for clinical evaluation and interpretation.

6.3.1 Programmatic evaluation of the results in Example use case

The resulting excel file can be seen in figures 2 and 3. Each target will have their own tabs presenting statistics for all group-target-pairs. Also a “combination” tab is provided to show the best findings from all significant group-target-pairs.



16	Lotta_Score_predictor ; 0/20p No risk in medication	62	B-Trom (2791)_at_least_one_value in_limits_in_insp_year_1Yes_0No_dich	2PASS	8,5651	8,5651	0,0776	0,0396	2,06
17	Lotta_4e_nausea ; 0: has not felt nausea in past 4 weeks	414	B-Hb (1552)_at_least_one_value above_limits_in_insp_year_1Yes_0No_dich	2PASS	0	0	0,0628	0,0393	2,06
18	Lotta_5_medication ; 0: it is not unclear how long medications will continue	275	B-Eryt (1341)_at_least_one_value above_limits_in_insp_year_1Yes_0No_dich	2PASS	0	0	0,0582	0,0004	3,53

Figure 2. Algorithm creates .xlxs file with individual tabs for all targets selected in data selection table (Table X) and a “combined” tab showing all targets.

To pass the custom test-schema and to get selected into the excel-file a group-target-pair needs to pass the following conditions :

- group count needs to be more than 10
- two tailed t-test p-value needs to be <0.001
- Bayesian factor needs to be more than 10

The group-target-pairs are then sorted based on :
1. TEST 2. abs_%-dif 3. count.

6.3.2 Understanding the results in Example use case

The process found over 100 relations. Only 3 of them is shown here with the permission from the head of AI research team. Most of the relations found are either “obvious” or “with no clinical relevance”. Usually this is the case for over 95% of the results. Said that, several interesting findings was found.

No significant relations was found for postal codes and only few weak relations with ATC-codes. Regarding ATC-codes, the n=518 is too small for such wide variety of medicines in use. More findings related to ATC-codes are expected in following years with comulative data.

The most important information in the excel file is listed below :

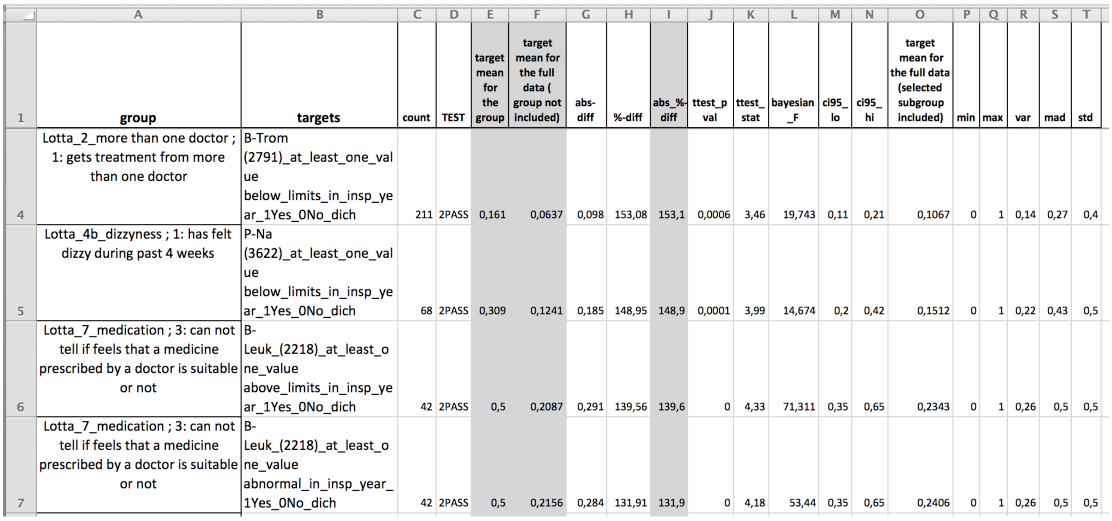
- **“group”**
    - The name of the group
        - In the row 4 in Figure 3 the group includes all participants who have done the LOTTA-questionary and answered “YES=1” to the question number 2 asking: “Have you been treated by more than one doctor during the last year?”. This means that all persons in this specific group have been treated by multiple doctors.
        - In the row 5 in Figure 3 it includes all participants who have done the LOTTA-questionary and answered “YES=1” to the question number 4b asking: “Have you felt dizzy during past 4 weeks”. This means that all persons in this specific groupmm have felt dizzy during past 4 weeks.
        - In the row 6 in Figure 3 it includes all participants who have done the LOTTA-questionary and answered “Can not tell”=3” to the question number 7 asking: “Have you felt dizzy during past 4 weeks”
- **“target”**
    - The target dimension for computing the mean against.
- **“count”**
    - The number of the participants belonging to the group (and also having any value in the target dimension)
- **“TEST”**
    - a custom test for evaluating the significance of the found relationship.
        - IF (ttest_pval <0.001) AND (bayesian_f >10) AND (count >10) THEN : 2Pass, ELSE 1Failed
        - This condition drops all comparisons from the results if the conditions are not met.
- **“abs_%_diff”**
    - absolute percentage difference between the “target mean for the group” and “target mean for the full data(group not included)

6.3.3 Indications of relationships in example use case

As can be seen in Figure 3 row 4, participants who get treatment from more than one doctor has significantly more values below the limits in B trombocytes (B-Trom(2791). 16,1% of the participants with more than one doctor have at least one value below limits whereas only 6,37% of the participants treated by just one doctor has values below limits. The difference is 153,2%. The group size is 211, p-value is 0.0006.

As can be seen in Figure 2 row 5, 30,9% of participants who have reported feeling dizzy during past 4 weeks have below normal values in sodium levels (P-na) whereas only 12,41% of the participant not reporting feeling dizzy have below normal values. The difference is 148,9%. The group size is 68, p-value is 0.0001.

The clinical interpretation regarding if these found indications of relationships are of any clinical or scientific value is not in the scope of this work and should be done by healthcare professionals after the relationships have been verified with e.g. regression tools.



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T
1	group	targets	count	TEST	target mean for the group	target mean for the full data (group not included)	abs-diff	%-diff	abs_%-diff	ttest_pval	ttest_stat	bayesian_F	ci95_lo	ci95_hi	target mean for the full data (selected subgroup included)	min	max	var	mad	std
4	Lotta_2_more than one doctor ; 1: gets treatment from more than one doctor	B-Trom (2791)_at_least_one_value below_limits_in_insp_year_1Yes_0No_dich	211	2PASS	0,161	0,0637	0,098	153,08	153,1	0,0006	3,46	19,743	0,11	0,21	0,1067	0	1	0,14	0,27	0,4
5	Lotta_4b_dizzyness ; 1: has felt dizzy during past 4 weeks	P-Na (3622)_at_least_one_value below_limits_in_insp_year_1Yes_0No_dich	68	2PASS	0,309	0,1241	0,185	148,95	148,9	0,0001	3,99	14,674	0,2	0,42	0,1512	0	1	0,22	0,43	0,5
6	Lotta_7_medication ; 3: can not tell if feels that a medicine prescribed by a doctor is suitable or not	B-Leuk_(2218)_at_least_one_value above_limits_in_insp_year_1Yes_0No_dich	42	2PASS	0,5	0,2087	0,291	139,56	139,6	0	4,33	71,311	0,35	0,65	0,2343	0	1	0,26	0,5	0,5
7	Lotta_7_medication ; 3: can not tell if feels that a medicine prescribed by a doctor is suitable or not	B-Leuk_(2218)_at_least_one_value abnormal_in_insp_year_1Yes_0No_dich	42	2PASS	0,5	0,2156	0,284	131,91	131,9	0	4,18	53,44	0,35	0,65	0,2406	0	1	0,26	0,5	0,5

Figure 3. Partial screen shot from “combined” tab of the .xlxs-output of the compare_categories_to_target()-function with settings defined in Table X.

7 DISCUSSION

Adequate dimension conversion and measurement scale transformation that doesn't loose information in the original data nor mistakingly introduce new information is mandatory for effective utilisation of modern DA and AI tools for processing medical data.

When data becomes high dimensional it is preliminary to develop methods for quickly selecting the correct dimension-transformations for different analysis and tools in a data-specific manner. The selection process as well as all programmatic processes should be possible to integrate to custom data-pipelines, makin the use of comericial tools inadequate. Also, in projects like Pori-75 where data is cumulative and grows in yearly basis, the adoption of the new data needs to be easy, fast and allow some annual changes in the data collection, recording and data struture.

New interesting data sources inspires the team to define plethora of pre-hoc hypotheses that should be quickly tested to identify the most potential hypothesis. Data-specific automated hypothesis testing is a good tool for such quick-and-dirty idea testing.

It is important to keep in mind that when comparing group means with automatic significance testing it is mandatory to stick to pre-hoc hypothesis and in traditional scientific research traditions.

Exploring trends from medical data by using subgroup analysis is very easily misleading. The use of group mean analysis in hands of people without deeper understanding is the reason for the partly unnecessary demonisation of subgroup analysis during the previous years.

On the other hand, the use of group mean analysis in hands of a professional can be a efficient, reliable and sometimes irreplacable tool. Justified arguments against using this tool is based on several misuses in scientific literature and underlines the need for deeper understanding when using such tools.

Arguments against pretty much any tool exists. The basic statistic tools having requirements in distribution or using variance have been misused for hundred years but has not been demonised as strongly as group mean analysis today.

Subgroup analysis conducted with good credibility criteria and strict testing methods for minimising the possibility of a change proportional to the size of the whole data - not just the group size - has always been and always will be a powerful tool for identifying potential relationships between dimensions.

REFERENCES

A. BRECK AND B. WAKAR. (2021). Methods, Challenges, and Best Practices for Conducting Subgroup Analysis. *OPRE REPORT*, *17*.

Anatomical Therapeutic Chemical (ATC) Classification. (2021). World Health Organization. https://www.who.int/tools/atc-ddd-toolkit/atc-classification

Dienes, Z., & Mclatchie, N. (2017). *Four reasons to prefer Bayesian analyses over significance testing*. https://doi.org/10.3758/s13423-017-1266-z

Gorard, S. (2005). REVISITING A 90-YEAR-OLD DEBATE: THE ADVANTAGES OF THE MEAN DEVIATION. *Source: British Journal of Educational Studies*, *53*(4), 417–430.

Guidelines for ATC classification and DDD assignment 2021. (n.d.).

Inglis, G., Archibald, D., Doi, L., Laird, Y., Malden, S., Marryat, L., McAteer, J., Pringle, J., & Frank, J. (2018). Credibility of subgroup analyses by socioeconomic status in public health intervention evaluations: An underappreciated problem? *SSM - Population Health*, *6*, 245. https://doi.org/10.1016/J.SSMPH.2018.09.010

Joshi, A., Kale, S., Chandel, S., & Pal, D. (2015). Likert Scale: Explored and Explained. *British Journal of Applied Science & Technology*, *7*(4), 396–403. https://doi.org/10.9734/BJAST/2015/14975

Kim, H.-Y. (2018). Statistical notes for clinical researchers: covariance and correlation. *Restorative Dentistry & Endodontics*, *43*(1), 4. https://doi.org/10.5395/RDE.2018.43.E4

Marateb, H. R., Mansourian, M., Adibi, P., & Farina, D. (2014). Manipulating measurement scales in medical statistical analysis and data mining: A review of methodologies. *Journal of Research in Medical Sciences : The Official Journal of Isfahan University of Medical Sciences*, *19*(1), 47. /pmc/articles/PMC3963323/

Holm, A., Kanninen, J.-C., Ampio, M., Teeri, S., Inberg, E., & Puustinen, J. (n.d.). Pori75-neuvolatoiminnan vaikuttavuus yksilö- ja väestötasolla. *Satakunnan Ammattikorkeakoulun Ja Seinäjoen Ammattikorkeakoulun Uuden Yhteisen Toimintamallin Satoa*.

Mishra, P., Pandey, C. M., Singh, U., Keshri, A., & Sabaretnam, M. (2019). Selection of Appropriate Statistical Methods for Data Analysis. *Annals of Cardiac Anaesthesia*, *22*(3), 297. https://doi.org/10.4103/ACA.ACA_248_18

Nestle Nutrition Institute. (n.d.). *Nutrition Screening as as A guide to completing the Mini Nutritional Assessment (MNA®) Screen and intervene. Nutrition can make a difference. 2 Mini Nutritional Assessment (MNA®).*

Phillips, P. (2009). Pitfalls in interpreting laboratory results. *Australian Prescriber*, *32*(2), 43–46. https://doi.org/10.18773/AUSTPRESCR.2009.022

Sintonen, H. (1995). *The 15-D Measure of Health Related Quality of Life. II Feasibility, Reliability and Validity of its Valuation System*.

Sleight, P. (2000). Debate: Subgroup analyses in clinical trials: fun to look at - but don't believe them! *Trials 2000 1:1*, *1*(1), 1–3. https://doi.org/10.1186/CVM-1-1-025

Upton, G., & Cook, I. (2008). A Dictionary of Statistics. *A Dictionary of Statistics*. https://doi.org/10.1093/ACREF/9780199541454.001.0001

Winter, J. F. C. de, & Dodou, D. (2019). Five-Point Likert Items: t test versus Mann-Whitney-Wilcoxon (*Addendum added October 2012*). *Practical Assessment, Research, and Evaluation*, *15*(1), 11. https://doi.org/https://doi.org/10.7275/bj1p-ts64