UNIVERSITEIT TWENTE.

Data Science [202300200]

Course year 2023/2024, Quarter 1A

DATE
October 5, 2023

EXCERPT

Projects

TEACHERS
Faizan Ahmed (TS)
Nacir Bouali (IENLP)
Faiza Bukhsh (PM)
Karin Groothuis-Oudshoorn (DM)
Maurice van Keulen (DEP,SEMI,PDBDQ)
Mahdi Khodadadzadeh (GIS)
Julia Mikhal (DM)
Estefania Talavera (CV&IC)
Shenghui Wang(DINT)

COURSE COORDINATOR
Maurice van Keulen (quartile 1A)
Maurice van Keulen (quartile 1B)
Faizan Ahmed (quartile 2A)

PROJECT OWNERS
Faiza Bukhsh
Karin Groothuis-Oudshoorn
Maurice van Keulen
Elena Mocanu
Mannes Poel
Michel van Putten
Mohsen Jafari Songhori
Luc Wismans
Mahdi Khodadadzadeh

Contents

Part II

Projects

Project 1

Project 1: Predicting surgical case durations for a Thorax centre [PSCD]

1.1 Introduction

Project owner: Karin Groothuis-Oudshoorn

Primary topic: DPV and/or DM

In modern healthcare, organizations face the challenge of delivering more and better quality care with less human and financial resources. This is mainly due to rising demand for healthcare and increasing expenditures. Efficiency is directly linked with quality, as inefficient care processes use up valuable resources and displace more useful care. Efficiency improvements are therefore very valuable for hospitals. MST is a top-clinical medical center located in the region of Twente and is one of the biggest non-academic hospitals in the Netherlands. The medical center compromises of two inpatient clinics in Enschede and Oldenzaal and two outpatient clinics in Haaksbergen and Losser. The inpatient clinic in Enschede has moved her patients as of 2016 to the newly built location Koningsplein. This new location provides a capacity of 739 beds. Thorax Centrum Twente (TCT) is a center within MST, specializing in diagnosis and treatment of cardiothoracic diseases. Multidisciplinary medical care is delivered through several cardiothoracic-related specialties such as cardiology and cardiac surgery. TCT is one of the 16 thorax centers in the Netherlands and has grown rapidly after its establishment in September 2004. One reason for this is their short waiting list for open heart surgery, making TCT an interesting medical center for patients. TCT performs approximately 1,100 to 1,200 open-heart surgeries per year, mainly coronary and heart valve surgeries. TCT experiences a high rate of operating rooms working beyond regular operating time. High amounts of overtime result in unnecessary costs and low staff satisfaction. A recent study among Dutch hospitals suggests that more accurate predictions of surgical case duration and altering the sequencing of surgical cases on an OR-schedule can improve efficiency[1].

References

[1] van Veen-Berkx E, Elkhuizen SG, van Logten S, et al. Enhancement opportunities in operating room utilization; with a statistical appendix. J. Surg Res. 2015;194(1):43-51.

Deliverables

- Report (according to the DS template)
- Presentation

1.2 Description of data set

The dataset comprised of 4087 surgical cases performed from January 2013 to January 2016 at TCT. The surgical case duration is given in minutes. The hospital stay time and IC stay time is given in days. Unknown data is indicated with 'NULL' or 'onbekend'. The data is in the file 'surgical_case_duration.csv'. In tables 1.1-1.4 you can find descriptions of the variables in the dataset. The different levels are given. In case of type of surgery not all labels are translated (we left out the less frequent ones).

1.3 Description of challenge

The challenge of this project is to identify patterns in surgical case durations and to derive prediction models and / or classification models for surgical case duration to support OR-planners at TCT in making the most efficient OR-schedules with the available patient level data in order to decrease the overtime at TCT, while maintaining the current OR-utilisation rate.

1.4 Tips and suggestions

- Some variables (features) have a lot of categories and some of those categories have few observations. In that case it can be wise to recode these categories to e.g. 'other types' or leave those observations out. But please explain and give arguments if you do so!

Table 1.1: Description of Surgery-related variables

Variablename	Variable (English)	Categories (definition)
Operatietype	Surgery type	Aortic Valve Replacement (AVR) AVR + MVP Bentall Procedure Coronary Artery Bypass Graft (CABG) CABG + AVR CABG + MVP Epicardial LV-lead (Epicardiale LV-lead) Lobectomy or segment resection (Lobectomie of segmentresectie) Mediastinoscopy Mitral Valve Plasty (MVP) MVP + Tricuspid Valve Plasty (TVP) Mitral Valve Replacement (MVR) Nuss bar removal Nuss-procedure Refixation of the sternum (Refixatie sternum) Rethoracotomy (Rethoractomie) Removal of steel wires (Staaldraden verwijderen) VATS Boxlaesie (video assisted thoracic surgery) Wound debridement (wondtoilet) Other types
Benadering	Surgical approach	Full sternotomy (Volledige sternotomie) Left antero lateral (Antero lateraal links) Right antero lateral (Antero lateraal rechts) Left postero lateral (Postero lateraal links) Right postero lateral (Postero lateraal rechts) Partial sternotomy (Partiele sternotomie) Other approaches: Parasternaal links, Parasternaal rechts, Dwarse sternotomie, Xiphoidaal, NULL)
Chirurg	Surgeon	Surgeon 1 – 15 Other specialism (Ander specialisme)
Anesthesioloog	Anesthesiologist	3 – 19, Unknown (onbekend)
OK	Operation room	HCK1, HCK3, HCK4, OK 1, OK 10, OK 11, OK 3, OK 4, OK 5, OK 9, TOK1, TOK2, TOK3, TOK4, else (Elders)
Casustype	Urgency	Elective (planned on the elective program) (Electief) Emergency ($<$ 24 hours) (Spoed $<$ 24 uur) Acute ($<$ 30 minutes) (Acuut $<$ 30 minuten) Acute (Spoed) Acute ($<$ 5 hours) (Spoed $<$ 5 uur) Unknown (NULL)
Dagdeel	Time of day	Morning (7:00 – 12:00) Afternoon (12:00 – 18:00) Evening and night (18:00 – 7:00)
Aantal anastomosen	Amount of bypasses	Continuous variable
HLM	Cardiopulmonary bypass use	Yes (heart-lungmachine usage planned for surgery) No

Table 1.2: Description of Surgery-related variables

Variablename	Variable (English)	Categories (definition)
Leeftijd	Patient age	Continuous
Geslacht	Patient gender	Male Female
AF	Presence of atrial fibrillation	Yes (AF rhythm present) No
Chronische longziekte	Presence of chronic lung disease	Yes (long term use of bronchodilators or steriods for lung disease) No
Extracardiale vaatpathie	Presence of extracardial arteriopathy	Yes (claudication, carotid occlusion or 50% stenosis, amputation for arterial disease or previous or planned intervention on the abdominal aorta, limb arteries or carotids) No
Active endocarditis	Presence of active endocarditis	Yes (patient still on antibiotic treatment for endocarditis) No
Hypertensie	Presence of hypertension	Yes No
Pulmonale hypertensie	Presence of pulmonary hypertension	Normal (no increased pulmonary artery pressure) Moderate (pulmonary artery systolic pressure 31-55 mmHg) Severe (pulmonary artery systolic pressure $>$ 60mmHg)
Slechte mobiliteit	Presence of poor mobility	Yes (severe impairment of mobility secondary to musculoskeletal or neurological dysfunction) No
Hypercholesterolemie	Presence of hypercholesterolemia	Yes No
Perifeer vaatlijden	Presence of peripherial vascular disease	Yes No
Linker ventrikel	Left ventricle	Good (LV ejection fraction $>$50%) (Goed) Moderate (LV ejection fraction 31-50% (Matig) Poor (LV ejection fraction $\leq$ 30% (Slecht) Very poor (Heel slecht)
Nierfunctie	Renal function	Normal (creatine clearance $>$ 85 ml/min Moderate (creatine clearance 50-85 ml/min) (Matig) Poor (creatine clearance $<$ 50 ml/min or dialysis) (Slecht) Dialyse
DM	Presence of diabetes mellitus requiring insulin	Yes (diagnosis DM requiring insulin) No
Eerdere hartchirurgie	Previous heart surgery	Yes (heart surgery in the patient's history) No
Kritische preoperatieve status	Critical pre-OR state	Yes (ventricular tachycardia or ventricular fibrillation or aborted sudden death, preoperative cardiac massage, preoperative ventilation before anesthetic room, preoperative inotropes or IABP, preoperative acute renal failure (anuria or oliguria $<$10ml/hr)) No
Myocard infact $<$90 dagen	Mycordial infarction before surgery	Yes (MI within 90 days before surgery) No
Aorta chirurgie	Aortic surgery	Yes (planned surgery on the aorta) No
Euroscore1	Euroscore1	Continuous variable
Euroscore2	Euroscore II	Continuous variable

Table 1.3: Description of Surgery-related variables (Continuation)

Variablename	Variable (English)	Categories (definition)
CCS	Canadian Cardiovascular Society (CCS) score for angina	0 (no symptoms) 1 (angina only during strenuous or prolonged physical activity) 2 (slight limitation, with angina only during vigorous physical activity) 3 (symptoms with everyday living activities, i.e. moderate limitation) 4 (inability to perform any activity without angina or angina at rest, i.e. severe limitation)
NYHA	New York Heart Association (NYHA) score - dyspnea	1 (cardiac disease, but no symptoms and no limitation in ordinary physical activity, e.g. no shortness of breath when walking, climbing stairs etc.) 2 (mild symptoms (mild shortness of breath and/or angina) and slight limitation during ordinary activity) 3 (marked limitation in activity due to symptoms, even during less-than-ordinary activity, e.g., walking short distances (20-100 meters). Comfortable only at rest 4 (severe limitations, experiences symptoms even while at rest, Mostly bedbound patients).

Table 1.4: Description of Outcome variables

Variablename	Variable (English)	Categories (definition)
Geplande operatieduur	Planned surgery duration	Continuous outcome
Operatieduur	Surgery duration	Continuous outcome
Ziekenhuis ligduur	Hospital days	Continuous outcome
IC ligduur	Intensive care days	Continuous outcome

Project 2

Project 2: Public Priorities for Primary Child Health Care for Children [MOCHA]

2.1 Introduction

Project owner: Karin Groothuis-Oudshoorn & Brenda Voorthuis

Primary topic: DPV

The health of our children is of utmost important not only for themselves and their families, but for the whole society. As future workers, parents and carers, these children are the ones that will build the world of the future. Health services for children are structured differently throughout the European Union, and there is little research into what works best. Therefore, the MOCHA project (Models of Child Health Appraised) performed a systematic, scientific evaluation of the types of primary care for children that exist in Europe (`http://www.childhealthservicemodels.eu/`). The aim of this project was to elicit formative values from the general public in five different countries and to determine public priorities in the assessment of the quality of a child-oriented primary health care system. The following research question and sub-questions were formulated. What are the priorities of European citizens in assessing the quality of primary care for children in Europe?

1. What are the experiences and/or perceptions of European citizens with the quality of currently provided primary care for children?
2. What are the preferences of European citizens with respect to the quality attributes of primary care for children?

Based on the diversity of the primary health care systems (general-practitioner-led, pediatrician-led or mixed), the United Kingdom, the Netherlands, Germany, Spain and Poland were chosen for studying these research questions. In a descriptive, cross-sectional, quantitative design, a uestionnaire was used to elicit preferences of a representative sample of the general public with respect to children's primary care and to measure experiences with the quality of currently provided care. Based on the review of literature, the child and carer centred outputs of each of the attributes of the primary health care system were efined. Through an iterative process within the MOCHA project team, the outputs were operationalized in 40 attributes-items. ttribute-items were operationalized in plain language and technical jargon was avoided as much as possible. Between one and nine attribute-items were operationalized for each of the nine attributes (outputs) of a child-oriented health care system (see Table 2.1).

Table 2.1: Description of the 9 attributes of the primary health care system

Attributes	Definition	Attribute-items
ACCESSIBLE	Accessible primary care is available within reasonable reach of parents and children, with ample opening hours, good appointment systems and other aspects of service organization and delivery that allow children to obtain the services when they need them"	1, 2, 3, 4, 5, 6, 7, 8, 9
AFFORDABLE	Affordable primary care can be accessed without inordinate financial barriers, such as high co-payments or cost-sharing arrangements	10, 11
APPROPRIATE	Appropriate primary care is effective in meeting the child's needs, timely and of high technical quality	12, 13, 14, 15, 16, 17
CONFIDENTIAL	Confidentiality in primary care is the right of a child to have personal, identifiable medical information kept private if they choose to, from medical professionals as well as parents	18, 19, 20
CONTINUOUS	Continuous primary care is the experience of a continuous caring relationship with the health care professional(s) by a single child and its parents over time, that is responsive of the child's changing needs.	21, 22, 23, 24, 25, 26
COORDINATED	Coordinated primary care is deliberately organizing child care activities and sharing of information among all of the participants concerned with a child's care with the aim to achieve safer and more effective care.	27, 28, 29, 30, 31
EMPOWERING	Empowerment in primary care is a process through which children and parents gain greater control over decisions and actions affecting a child's health	32, 33, 34, 35, 36, 37
EQUABLE	Equable primary care is the absence of systematic and potentially remediable differences in access to primary care and health status across population groups"	38, 39
TRANSPARENT	Transparent primary care is the degree to which a healthcare service or provider is open to children and parents about their quality, cost structure, services and work method	40

The description of the attribute items can be found in the file **mocha_items.csv**.

Deliverables

- Report (according to the DS template)
- Presentation

2.2 Description of data set

Two data files are provided containing the same data. In the file **POCHA_FULL_DATASET_names.xlsx** the column names are the variable names and the labels are the values. In the file **POCHA_FULL_DATASET_labels.xlsx** the column names are the labels of the variable names, i.e. the question itself. The values in the cells are not the labels but the values. The file **mocha_items.csv** contains the description of the attribute items.

2.2.1 The POCHA questionnaire

The POCHA questionnaire consisted of five parts (see : `https://www.childhealthservicemodels.eu/wp-content/uploads/Final-Report-POCHA.pdf`).:

1 background characteristics

2 the health status of the child(ren) and current health care consumption of any child(ren) below the age of 18

3 child independence, i.e. the age at which a child could be or should be able to make decisions independent from its parents

4 the quality of the primary health care system

5 the prioritization of the 40 attribute-items of a child-oriented primary health care system

The following background characteristics of respondents were measured:

- age: 19 years or younger, between 20 and 24, . . . , between 65 and 69, 70 years or older;
- gender: female, male;
- country: United Kingdom, Netherlands, Germany, Spain, Poland; and the countries' region;
- number of children: 1, 2, 3 other;
- number of children ¡ 18: yes/no;
- highest level of completed education: with the following categories for the UK: Entry level, GCSE (grades D-G), GCSE (grades A*-C), A-Level, Higher National Certificate/National Diploma, Bachelor Degree, Master Degree, Doctoral Degree
- size of the city, where the respondent lives: ¡ 100, 100 – 1.000, 1.000 – 10.000, 10.000 – 20.000, 20.000 – 100.000, 100.000 – 200.000, 200.000 – 1.000.000, ¿ 1.000.000.

Health Status and Health Care Consumption

Health status was measured was asking: "In the past 12 months, has (one of) your child(ren) had a medical condition that lasted longer than 6 weeks?". "What condition was this/were these? Please indicate all conditions that apply for any child(ren) below the age of 18." Conditions to choose were: Eczema, Asthma, Hay fever, Allergy, Stomach ache, Headache, Back problems, Fatigue, Sleep problems, Depressive complaints, Hyperactivity and ADHD, Constipation, Overweight and obesity, Other: . . . .

Following the question on health status, health care consumption for children ¡ 18 with long-term disease (¿ 6 weeks) in the previous 12 months was measured with the frequency of contact with primary and secondary health care providers.

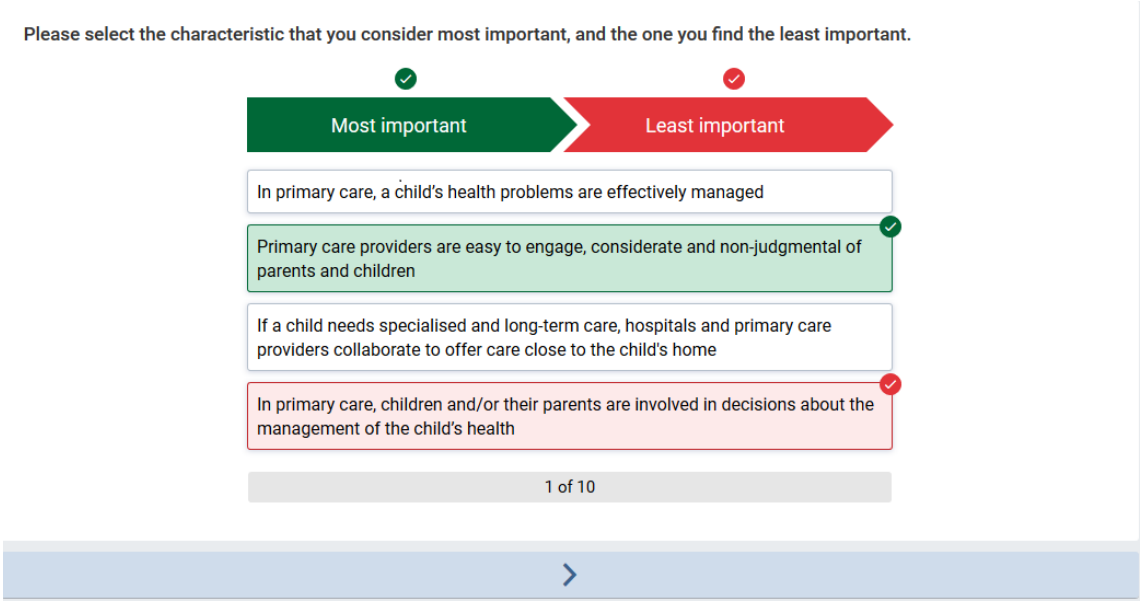

Figure 2.1: Example of a prioritization question from the English Survey.

Quality of the Primary Health Care System

In order to measure experiences with and/or perceptions of primary care for children, ten questions with statements were presented about the quality of primary health care system for children in place in the respondent's country. The respondent was asked to indicate to what extent he agreed with each statement, on a 5-point rating scale (strongly disagree to strongly agree). If the respondent had no direct experience of primary care for children, he was asked to react on the statement based on his perceptions of primary care for children in his country. These perceptions could be based on media coverage and/or on stories from friends and family. For each respondent on the questionnaire, the ten statements were randomly selected out of 40 statements on the quality of the primary health care system for children.

Prioritization of Attribute-items of a Child Oriented Primary Health Care System

To determine what people find important characteristics (= attribute-items) of the quality of a primary health care system, we asked respondents in 20 questions to indicate what they think is important in the assessment of primary care for children. To this respect, best-worst scaling questions were used; for an example of this type of question used, see Figure 1. In each question, a list of four potential characteristics of primary care was presented. The respondent is asked to read these characteristics carefully and to choose which of these he finds most important in primary care for children and which of these he finds least important.

Respondent priorities in each question are likely influenced by the combination of attribute-items in that question. Thus, combinations of attribute-items within and over respondents was varied using an experimental design. Each attribute-item was presented to each respondent twice, in a combination different from each other. Eight different sets of ten questions were made, in which each attribute-item was presented once. Two different sets of ten questions were presented to each respondent.

2.3 Description of challenge

The challenge in this project is to provide an appropriate star schema for the data that is obtained from the questionnaire and to summarize and visualize the data such that the research questions are answered. Look for relevant differences between countries and based on other background characteristics.

2.4 Tips and suggestions

- The best-worst scaling questions can be analyzed with different methods, where the most easy methods is the best minus worst counts analysis. As an example you can look at the article from Louviere & Flynn "Using Best-Worst Scaling Choice Experiments to Measure Public Perceptions and Preferences for Healthcare Reform in Australia, the Patient 2010". Priorities for attribute-items of quality of care can be calculated using counts analysis on an individual and a group level. First, the number of times an attribute-item of quality was selected as most (max 2) and least important (max 2) was counted per respondent. Then, best-worst scores were calculated by subtracting the number of times each attribute-item was selected as least important from the number of times it was selected as most important. Individual best worst counts ranged from – 2 (not important) to +2 (important). Group priorities of attribute-items can then be defined as the best-worst score of that attribute-item for that group (defined e.g. by country). Best worst scores should then be normalized over groups by dividing the best-worst count by the number of times each attribute-item was presented to the group (2 times for each respondent in the group) and multiplying this ratio with 100. Group best worst scores can be calculated for respondents from each country, male and female respondents and respondents with and without children below the age of 18.

Project 3

Project 3: Predicting mortality for COVID-19 patients [COVID]

3.1 Introduction

Project owner: Brenda Voorthuis & Karin Groothuis-Oudshoorn

Primary topic: DM

The COVID-19 outbreak was first identified in Wuhan, China, in December 2019 and was declared a global pandemic in March 2020. By the end of July 2020, there were over 17 million confirmed cases worldwide and close to 700,000 deaths. The risk of death increases with age and with underlying conditions such as obesity and diabetes. The worldwide increase in COVID-19 cases is putting high pressure on healthcare services. Early assessment of the severity of COVID-19 cases is essential for logistic planning. Studies by Yan et al. (2020) and Zhou, Chen, and Lei (2020) suggest that the mortality of an individual COVID-19 case can be predicted with more than 90% accuracy and over 10 days in advance based on just three biomarkers: lactic dehydrogenase (LDH), lymphocyte, and high-sensitivity C-reactive protein (hs-CRP).

References

Yan, L., Zhang, H-T., Goncalves, J., Xiao, Y., Wang, M., Guo, Y., ... Yuan, Y. (2020). An interpretable mortality prediction model for COVID-19 patients. *Nature Machine Intelligence, 2*, 283-288. doi:10.1038/s42256-020-0180-7

Zhou, F., Chen, T., & Lei, B. (2020). *Do not forget interaction: Predicting fatality of COVID-19 patients using logistic regression*. Retrieved from `https://arxiv.org/abs/2006.16942`

Deliverables

- Report (according to the DS template)
- Presentation

3.2 Description of data set

Two datasets are available. The first dataset was used as training data in Yan et al. (2020) and was collected between January 10, 2020 and February 18, 2020, at the Tongji Hospital in Wuhan, China. It consists of 375 patients, of which multiple biomarker measurements were taken and 174 died. The second dataset of another 110 patients from the same hospital was used as test data and was collected between February 19, 2020, and February 24, 2020. Of these 110 patients, 13 died.

The training data file **time_series_375_preprocess_en.xlsx** contains data on the age and gender of each patient, along with admission time, the outcome (death/discharge) and time of discharge/death. The other 74 columns contain information on different biomarkers. The test dataset, **time_series_test_110_preprocess_en.xlsx**, contains the same patient information but only the biomarkers LDH, lymphocytes, and hs-CRP.

3.3 Description of challenge

The challenge of this project is to develop a prediction model for the mortality of COVID-19 patients based on the biomarkers that are available in the data. Create multiple machine learning models (i.e. different DM methods) and compare them based on the applicable performance measures to identify the best fitting model.

3.4 Tips and suggestions

- Multiple measurements per patient are available in the datasets. In the development and testing of the models by Yang et al. (2020) and Zhou, Chen, and Lei (2020), the last known sample of each patient was used. Other methods are allowed, but explain and substantiate your choice if you do so!
- The model output corresponds to mortality. The variable `outcome` is coded with 0 (discharge) and 1 (death).
- The test dataset only contains information on the three biomarkers LDH, lymphocytes, and hs-CRP. If you want to test a model containing different biomarkers, split up the training data into another training dataset and testing dataset or use cross-validation.