MSc Mathematics

Master thesis

Pattern recognition and feature engineering in mental health diagnostics

by

Olga Półchłopek

August 24, 2018

Supervisor: dr. Mark Hoogendoorn

Second examiner: prof. dr. Sandjai Bhulai

Department of Mathematics
Faculty of Sciences

VU UNIVERSITY AMSTERDAM

Abstract

Mental health problems are widely underdiagnosed and have a negative impact on everyday life. Prevalence rates for such disorders in Dutch children and adolescents vary between 7% and 30%. A retrospective study examines Electronic Medical Records from general practices in the Leiden area to develop a tool to identify young patients at risk of suffering from mental disorders. Feature engineering methods tailored for clinical data are proposed together with a mathematical setup and a modified algorithm for mining temporal patterns. A new pattern-scoring scheme based on the Wilson score interval is provided to obtain frequent and predictive patterns as well as to accelerate the mining process. Six machine learning models (logit, SVM, regression tree, random forest, deep neural network, XGBoost) are trained on five age groups achieving AUC values of 0.75-0.79 with sensitivity and specificity above 0.7.

The study is a part of a project called "Primary care integrated for identification of psychosocial problems in children" conducted in the Department of Public Health and Primary Care of Leiden University Medical Center.

Title: Pattern recognition and feature engineering in mental health diagnostics
Author: Olga Półchłopek, olga.polchlopek@gmail.com, 2598758
Supervisor: dr. Mark Hoogendoorn
Second examiner: prof. dr. Sandjai Bhulai
Date: August 24, 2018

Department of Mathematics
VU University Amsterdam
de Boelelaan 1081, 1081 HV Amsterdam
`http://www.math.vu.nl/`

Contents

1. Introduction

Mental health problems have been receiving more and more attention with the evolution of social media [62] and due to their soaring rates in new generations. Such disorders have a negative impact on everyday life and might lead to repercussions in wellbeing and functioning, especially if experienced in childhood [24, 26, 37, 48, 56]. A rarely acknowledged fact is that they are relatively common in children and teenagers [36, 61], with prevalence rates given by the literature varying between 7% and 30%[1] [12, 29, 30, 38, 70] and the estimated scope of the problem, including unidentified cases, reaching 50% [105]. Mental health issues are widely underdiagnosed in children. A substantial number of young patients is not recognized as having any problems of the kind [22, 72, 80, 110], as primary care professionals fail to identify 43-50% of cases with elevated scores on mental health screening tools [12, 102]. As a consequence, mental health remains insufficiently treated, with a large proportion of children in need not receiving adequate help [79, 89].

The scope of the problem has bean soaring. Among teenagers, rates of depression and anxiety have increased by 70% in the past 25 years, particularly since the 1980's, and evidence for a rise in emotional problems has been shown [18]. The number of students disclosing a mental illness when they arrive at university has grown fivefold in the past decade [88]. Safer [78] reported consistent increases in diagnostic prevalence of ADHD, whereas [91] noted that suicides more than doubled among women and grew by 40% in men aged 13-18 between 2007 and 2015. Many sources emphasize the role of online social networking sites in exacerbating mental health status in children and young adults, causing eating and sleeping disorders to spread gradually [46, 85, 86]. Since a positive correlation of time spent using social media and depression has been found [63], there might be a threat of mental health problems increasing in the future.

This study is a part of a project called "Primary care integrated for identification of psychosocial problems in children", in short PIPPI, conducted at the Department of Public Health and Primary Care of Leiden University Medical Center. It is a retrospective research with healthcare data from children enlisted with general practice centres in the Leiden area. The objective is to develop a tool for proactive identification of children and adolescents at risk for mental health problems in primary care by analyzing Electronic Medical Records (EMR) from general practitioners (GP-s) and child health professionals (CHP-s). In line with the project, the aim of this study is focused on extracting information from a particular dataset more than extensively testing the presented methods.

It is believed that GP-s have a central role in identification, treatment and referral of mental health problems in children [95]. They are able to maintain a long-lasting

[1] A plausible explanation for such a high variance is given in section 2.3.

relationship with their patients, monitor changes in the family system over time and conclude the possible effects on the child's health [101, 110]. This professional acquaintance with a child has been proved to have a beneficial influence on mental issues recognition [79, 111]. A majority of children and adolescents in the Netherlands visit their GP at least once a year [44, 92] and primary care attendance rates do not vary among different healthcare systems [7, 95, 101], which assures frequent screening opportunities regardless of the country. It is therefore perfectly justified for GP-s to be potential recipients and users of the decision supporting tool, especially that training and employment of protocols can reduce inter-professional variation in the short term according to some studies [72, 100]. The aim of developing highly predictive models is not only to help the doctors diagnose mental disorders, but also to predict and counteract where applicable.

Historical information is commonly used in diagnostics. Decision support systems are able to operate on big datasets and reason quicker than human brains, hence the increasing interest in such tools. The earliest prototypes were MYCIN system for choosing antibiotics [17], deDombal for diagnosing abdominal pain [58] and HELP that provided medical alerts [99]. Since then, such systems contributed to development of widely-used methods [53]: inductive symbolic learning (top-down induction of decision trees, decision rules and induction of logic programs) such as Assistant-R [14, 41] or LFC [68], statistical or pattern recognition methods (k-nearest neighbours, discriminant analysis, and Bayesian classifiers) such as the Semi-Naive Bayesian Classifier [40] and neural networks (multilayered feedforward neural network with backpropagation learning, the Kohonens self-organizing network and the Hopfields associative memory), e.g. Backpropagation with weight elimination [77].

In the field of mental health problems early research showed little evidence of success. In 1996 Lewis [47] created a computerized self-assessment system of common mental disorders (PROQSY) which showed short-term improvement compared to lack of such assessment. No difference was reported in long term. CaseWalker by Cannon [13] was a reminder system based on a list of rules screening for mood disorders. It proved that computer alerts were more effective in documenting guideline criteria for major depressive disorder than manual actions. In Schriger's PRIME-MD referral system [82] patients completed additional forms that frequently indicated a need of a psychiatric treatment. However, medical experts did not adhere to the system's referrals, diagnosing the same rates of mental health problems in the referred and control groups. In 2002 Rollman [75] revealed that screening for major depression, electronically informing GP-s of the diagnosis, and then exposing them to evidence-based treatment recommendations, has little differential impact on clinical outcomes. These decision systems did not adopt advanced machine learning techniques but were connected to EMR.

Other attempts applied decision trees [10], constrained logic programming (if-then clauses) [107], Brain Imagining [43] and Fuzzy Logic [76]. Support Vector Machine, Bayesian Network, Logistic Regression, Radial-Basis Function, Random Forest and Polygenic Scoring were used to infer mood disorders from genome data [52]. A more recent study [2] used features such as attention arousal, behavioural problems or CBCL score (a checklist to identify problems in children) to predict mental health problems. It then compared three methods, Multilayer Perceptron, Multiclass Classifier and LAD Tree,

achieving satisfactory results[2]. In this study some state-of-the-art techniques mentioned above (i.a. decision trees and neural networks) are used to build models based on non-psychosocial data with the strongest focus on the preprocessing, which is a crucial step in learning from EMR. No changes are proposed to the learning methods.

EMR allow for retrieving information in a temporal manner, i.e. preserving the chronology of the symptoms raised by patients. The order of medical history will be included as patterns, concatenating two or more consecutive and co-occurring events. A special framework for defining and mining the patterns needs to be designed to account for the complexity of EMR and present the data possibly unbiased. A study has shown that experts among medical doctors are more likely to employ pattern recognition in comparison to their non-expert peers [49]. This is especially important, as it would be intuitive for the GP-s to use the tool if they were familiar with the general idea behind the algorithm. From the economical point of view, which may be of interest for the central healthcare provider, the tool creates an opportunity to optimize the cost-efficiency of mental health treatment.

Factors known as the most important or the most studied will be included in all models and the results compared with medical literature if the models indicate them as significant. For instance, children's age was documented to have a positive [9, 71, 81], negative [12, 97] or no relation with mental health problems [51, 73]. Sex turns out important in school-aged kids, with boys being more inclined to experience mental and behavioural disorders [29, 45, 97, 110]. No association with gender has been proved for adolescents [12, 51, 73, 110] or toddlers [38, 71]. Another studied indicator is a number of visits at a GP [73]. These determinants, together with other features derived from EMR, will be evaluated both in linear an non-linear models.

The following chapter delves into description of the dataset and displays the intricacies posed by EMR. Chapter 3 has two main parts contributing to modeling mental health issues. The first one illustrates feature engineering techniques addressing the aforementioned complications and the other provides definitions, together with a mathematical setup, for mining patterns. The framework builds on the approach in [4], yet adapts it to the available data and extends it with a tractable method of scoring. The experimental setup can also be found in this chapter. An empirical application of these methods, serving as an attempt of developing models for the PIPPI project, is presented in chapter 4. The algorithms succeeded in meeting the requirements of the project. Both temporal and non-temporal data was used to train the models, in order to measure the impact of patterns. Further research directions are discussed in chapter 5 and a few useful algorithms coded in python are shared in the appendix.

[2]AUC of 0.88, 0.9 and 0.78 respectively. For AUC evaluation description refer to section 3.4.

2. Data

Healthcare industry digitalizes and stores big amounts of data referred to as Electronic Medical Records (EMR), arriving with complex and multivariate temporal datasets [4]. Each visit is registered in a system together with symptoms, life events and concerns raised by the patient or observed by the doctor that create multiple time series of clinical variables. This is kept mostly for recollecting medical history during the visits that follow, as well as statistical and reporting purposes. However, with the evolution of data science, the demand has risen for decision support systems which operate on big datasets and are able to reason quicker than human brain. It is a very natural step, as historical information is commonly used in diagnostics to compare and find similarities in medical cases that are obscure or rare, and therefore uneasy to solve.

2.1. Description

The data used in this study was collected for the PIPPI project [39]. It is a collection of EMR from 76 general practices in the Leiden area, gathered, concatenated and preliminarily aggregated by a third party[1]. It consists of records of all children of ages 0-19 on the 1st of January 2014 and still registered in a GP practice on the 1st of January 2015. Theoretically, all underage inhabitants are supposed to be registered at a GP. For those patients all available medical history is used until the end of 2017.

The collection consists of 6 tables, *patients, episodes, consultations, medication, tests* and *referrals*, that can be combined using a key variable which is a pseudonymized patient identifier. For the models, information about the birth year and sex of the patients is used, together with dates of their entering and exiting the system that are recorded in *patients*. The *episodes* table contains ICPC codes (International Classification for Primary Care) which are a WHO-approved way to record symptoms and events exemplified in table 2.1 and fully listed in [115]. The codes were aggregated into episodes that ideally should reflect all problems a patient has had, independently of the number of visits paid to their GP regarding those problems. On the other hand, *consultations* consists of records per visit. There are 1.46 mln distinct visits recorded.

For example, a patient can have multiple encounters with their GP on different dates, regarding contraception, coded `W12`. They would see the doctor first to discuss the contraceptive options, during the next visit an IUD (intrauterine contraceptive device) would be placed and the third appointment would be a scheduled check-up 6 weeks after the placement. These events could result in 1 row in the *episode* file under `W12` and

[1]Stichting Informatievoorziening voor Zorg en Onderzoek (STIZON)

3 such rows in the *consultations* table. If the patient had the IUD placed twice, this may appear as two separate events. However, there is no reference or key between the two tables and it depends on the doctor whether they code the event once or twice. No assumptions can therefore be made about the relationship between these sources, yet it could be a direction towards improving transparency of the dataset, and perhaps performance of research projects using it, to investigate this dependence.

Table 2.1.: Examples of ICPC codes by chapter

chapter	description	example
A	General and unspecified	A03 - Fever
B	Blood, blood forming organs, lymphatics, spleen	B73 - Leukemia
D	Digestive	D06 - Other abdominal pain
F	Eye	F03 - Teary eye
H	Ear	H01 - Earache
K	Circulatory	K84 - Cardiomyopathy
L	Musculoskeletal	L04 - Chest symptoms / complaints
N	Neurological	N79 - Concussion
P	Psychological	P72 - Schizophrenia
R	Respiratory	R74 - Acute upper respiratory tract infection
S	Skin	S18 - Crack / cut
T	Endocrine, metabolic and nutritional	T90 - Diabetes
U	Urology	U71 - Cystitis / urinary tract infection
W	Pregnancy, childbirth, family planning	W10 - Morning after pill / postcoital contraception
X	Female genital system and breast	X71 - Gonorrhea (female)
Y	Male genital system	Y05 - Swollen testicles
Z	Social problems	Z23 - Loss / death of parents / family

Source: Nationaal ICT Instituut in de Zorg[2][115]

The episode list dates back the furthest of all tables, it is thus an important overview of childhood affairs. The other tables were initialized 10 years ago. The records that are not ICPC-coded contain binary responses for descriptive symptoms text mined from the GP's notes. Such features include *stress, behavioural problems, ADHD, autism* and *sleeping disorders*. From the remaining datasets, variables for medication, tests and referrals to specialists are extracted. ATC stands for Anatomical Therapeutic Chemical and is another WHO certified system for identifying specific drugs or substances in prescribed medication [114]. A full code consists of 7 characters, each narrowing down

[2] https://decor.nictiz.nl/ketenzorg/kz-html-20141013T173536/index.html

the category of medication [116]. The categories are based on function and composition (see table 2.2).

Table 2.2.: ATC code groups

prefix	**group name** and example
A	**anatomical main group** Alimentary tract and metabolism
A10	**therapeutic subgroup** Drugs used in diabetes
A10B	**pharmacological subgroup** Blood glucose lowering drugs, excl. insulins
A10BA	**chemical subgroup** Biguanides
A10BA02	**chemical substance** metformin

Source: WHO[3]

2.2. Sanitization

All exclusion was made in agreement with medical experts due to the missing data (e.g. system entry date) or insupportable incoherence (such as when the calculated *age* or *exposure*[4] were negative numbers or the sex was ambiguously coded). Table 2.3 portrays the pipeline for eliminating those outliers with reasons and the respective numbers of patients cut from the study. It is worth noticing that the patients were removed in order and hence out of the 13 records with the sex coded wrong 2 were deleted due to a different reason.

Table 2.3.: Patients deleted from the dataset

102250	Initial number of patients
- 172	Deleted due to age < 0
- 705	Deleted due to the lack of entry date
- 7279	Deleted due to exposure < 0
- 11	Deleted due to unknown gender
94083	**Remaining patients**

Source: own calculation

2.3. Issues

To fully understand the reasoning behind the methods proposed in this paper, it is vital to acknowledge all the challenges that emerge when processing EMR. Some problems

[3]https://www.whocc.no/atc/structure_and_principles

[4]See definitions 3.2 and 3.3 in section 3.2.

frequently occur in other datasets. These consist of sparsity, homogeneity, imbalance and existence of multivariate attributes explained in more detail in the paragraphs that follow. However, there are many that require concepts tailored specifically for clinical data and mental health diagnostics – differing EMR systems and rules applied across general practices, uneven time gaps between events or inability to compare test results.

The EMR data inflicts trouble starting with its concatenation, as the registry systems are not uniform among healthcare units and the doctors may have different ways of recording events. For example, they can record only the main purpose of a visit or all disturbances mentioned by a patient. Chronic diseases might be reported at all times or only upon recognition, with each cyclical prescription registered or one bulk record of having certain medication prescribed. Therefore it is impossible, without detailed knowledge of the rules each doctor applies, to eliminate bias in studies based on the EMR derived from multiple general practices. A similar kind of distortion can be introduced by patients themselves, as it is only what they choose to present that can be reported by a GP in the system. It is thus a difficult but crucial task to arrive with consistent logic with well-justified assumptions that minimize this bias.

The complexity of the EMR follows mostly from the fact that it is multivariate, with hundreds of incomparable levels (their distance cannot be measured with a metric other than discrete and they resist being ordered from worst to best)[4]. A common approach for dealing with such data is one-hot encoding, i.e. spreading one factor column into many binary ones, each standing for one level of that factor [25]. However, with EMR data it creates a very sparse matrix (containing a lot of zeroes), so one has to be careful with quantitative methods, as they might work faultily. On the qualitative side, expert knowledge can be used to infer dependency between those levels as shown in section 3.2.3 for grouping medication.

Homogeneity of the data is another idiosyncrasy. Some clinical variables, e.g. laboratory results, can have numerical levels (counts, densities, time intervals), ordered (type 1, type 2, etc.) or categorical ones (positive, negative). This prevents applying uniform methods to all features and therefore prolongs the preprocessing [69]. Each executed transformation has to be justified from medical and mathematical point of view which very often enforces compromise on both sides. This issue is dealt with in section 3.2.2 and in mining algorithm construction in 3.3.5.

Health-related data is challenging also due to the underrepresentation of the target, which results in highly imbalanced datasets. Such data should be treated differently, as Machine Learning methods tend to create many false positives [50]. The usual approach would be to over- or undersample, [67] or to generate artificial records, which is a non-trivial task when dealing with sparse data. In the case of this research the data is only slightly unbalanced, with the target patients (i.e. patients who are recognized as having mental problems[5]) constituting to 27% of the population. No balancing methods were used, although it could be a possible step to explore further in the PIPPI project.

There are two reasons why the underrepresentation exists. Foremost, it is because mental disorders are underdiagnosed [104] – if a GP does not recognize and record in

[5]Refer to section 3.1 for a precise target definition.

the system that a patient is psychologically concerned, this patient will be considered a non-target instance. WHO revealed that GP-s identified only 49% of mental disorders confirmed by the study and furthermore that merely a half of the recognized cases received any help. [93] The second reason is that the patients themselves do not discuss their mental state with doctors. According to the study of International Consortium of Psychiatric Epidemiology [8], 74.6% of respondents in the Netherlands who received mental health treatment consulted their GP. [105] claims that 50% of the population will suffer from at least one mental disorder in their lifetime and that 25% have experienced such a problem in the past year. On the other hand, after [109], 6.6% of Dutch children and 7.5% of adolescents suffer from mental health problems, which would suggest that from a clinical perspective the target of 27% is an exaggerated estimate. The discrepancy results mainly from different designs of the studies. [109] counts only the young patients who were referred for mental health services during the 12 months preceding the assessment, whereas this research investigates issues in the entire medical history. The choice of respondents is almost as important – the first study used random selection of all inhabitants of certain age and the latter focuses only on those who visited general practices. For these reasons the share of patients concerned with mental disorders is expected to be closer to the one mentioned by [105] but not as high due to the aforementioned aversion to seek advice on mental health.

Time series approaches are not to be applied as the EMR events cannot be treated as such. This hardship is caused by the irregularity of visits to GP-s – there is no uniform time unit to be defined, as a patient can see their doctor twice in one week and then not make another appointment for a year. Moreover, the visits tend to be infrequent, so there are usually multiple problems that occurred between them. It is the doctors who decide how many of the issues should be recorded in the system. The outcome of this registration is a list of events with the date of the visit during which they were mentioned. The order or co-occurrence of such events cannot be inferred from the data, neither can their duration be assumed. The approaches this paper presents include introducing exposure in definition 3.3 and modifying patterns to capture co-occurrence in definition 3.10.

Explicitly in this case, a big obstacle is the lack of reference for test results. This is mainly because the study is dealing with children's health and the ranges for outcomes considered good are dependent on age and laboratory used, it is therefore difficult to create respective bounds [98]. This is less of an issue with adult patients. As a consequence, it is impossible to use the numerical values of the test results in the models and thus substantial information is lost. Furthermore, outcomes of binary response tests (positive/negative) are coded differently depending on the GP and the system they use. Examples of such coding are: `pos`, `POS`, `ONOOIT`, `0HA`, `0JA`, `volgt`, `0SPEC`, `N.AANT.`, `n.aant.`. Such strings can neither be compared nor one-hot encoded. An attempt to overcome this hardship was made in section 3.2.2.

As mentioned in 2.1, there exists no link between episodes and ICPC-coded events in the *consultations*. It may happen that what should be two separate episodes is recorded as one, with twice as many visits in the *consultations*, there are also events in the *consultations* missing an aggregating episode record and vice versa. Considering this,

these sources must be treated as separate. An unbiased approach would reject one of them to avoid double counting of some events, yet lose a lot of information at the same time. A solution is proposed in section 3.2.1.

3. Methodology

It is challenging to apply standard Machine Learning tools in modeling EMR data as mentioned in section 2.3. Tracing chronic mental disorders, as opposed to predicting acute diseases, requires a thorough understanding of the matter and using unconventional approaches. Mental health cannot be measured or compared like the output from intensive care unit electronic equipment – the lack of numerical variables poses the main inconvenience – hence the range of available instruments is scarce and the existent methods are more time- and memory-consuming than using numerical columns. The first two sections describe processing the EMR data in this particular case and serve mainly to explain the rationale behind all transformations, rather than introduce groundbreaking techniques. Nevertheless the process should not be depreciated, as the quality of features impacts the results more than any model tuning. The third section builds upon Batal's approach in [4] and adapts temporal patterns to mental health diagnostics. Similarities include definitions and the choice of relation but both the means of reducing the number of patterns and mining procedure must be revisited. Such adjustments are necessary, as with this type of data Batal's assumptions – a fixed-width observable time window in medical history and a lack of overlapping events – do not hold. Section 3.3.4 is a proposition of an improvement to the approach that can be used for other purposes as well. It provides a different method of selecting the most frequent and predictive patterns.

3.1. Target identification

The PIPPI project defines three simultaneous sources to determine if a patient was diagnosed with a mental health concern: ICPC codes, prescribed ATC-s and referrals to specialists who deal with psychological problems. Table 3.1 provides the codes used to identify those patients and the dates the diagnoses were made. The ICPC-s starting with `P-` stand for psychological disorders, `T06` is for anorexia/bulimia, the distinguished medication are psycholeptics (`N05`), psychoanaleptics (`N06`) and other nervous system drugs (`N07`). The codes given for ATC that are shorter than 7 characters define an entire category of drugs, e.g. `N07BB` stands for *Drugs used in alcohol dependence* which refers to 5 different substances: disulfiram, calcium carbimide, acamprosate, naltrexone, nalmefene. Appearance of any of them is used as a target identifier.

Table 3.1.: Values of features that identify target patients

source	value
ICPC	P1-30, P70-99, T06, T06.01, T06.02
ATC	N06BA02, N05A, N05B, N05C, N06A, N06BA04, N07BA, N06BA09, N07BB
referral	eerste-lijnspsychologie, EERSTE-LIJNSPSYCHOLOGIE, GGZ-instelling, psychiatrie, PSYCHIATRIE, psychologische zorg, PSYCHOLOGISCHE ZORG, psychotherapie, PSYCHOTHERAPIE, ELP, ELP eerste-lijnspsyc, ggz, GGZ, PSL, PSL psychologische z, PSL Psycholoog, PST, PSY psychiatrie, PSY, Psychiatrie, PTH, PTH psychotherapie

Source: PIPPI [39]

Since the end product of the PIPPI project is a software tool for recognizing children and adolescents at risk, the study is focused on observing patients before being diagnosed and modeling the probability of developing a mental illness in the future (e.g. a year or two). It is therefore important to exclude the medical history of target patients within a fixed time window before the first record containing any of the events from table 3.1 and all the post-recognition history of the patient. The time frame suggested by the medical experts is 180 days. Removing these records additionally prevents the target features to be included as explanatory variables.

3.2. Feature engineering

In the *patients* table the columns of interest are *birth year*, *entry date* and *exit date*, the last two being the dates of registering and unregistering from a general practice. Patients with missing values for birth and entry are removed from the study (see table 2.3) and for patients that have never exited the system, a new exit date must be defined.

Definition 3.1. Let p_0 be a non-target patient and $exit\ date(p_0)$ be the date of their signing out from the system.
If $exit\ date(p_0)$ is NULL (p_0 never unregistered or didn't unregister before 2018), then $exit(p_0) = 2017$.
If $exit\ date(p_0)$ is not NULL, then $exit(p_0) = \text{year}\,(exit\ date(p_0))$
Let p_1 be a target patient and $exit(p_1) = \text{year}\,(diagnose(p_1) - 180)$ where $diagnose(p_1)$ is the date of the first record of p_1 containing any of the values in table 3.1.

In other words, we can define *exit* as the year when we stop observing the patient. This is important for calculating their age at the end of their medical history, which is precisely the moment in which we want to predict their vulnerability to mental health disorders. It is an equivalence of what GP-s have at their disposal when they see their patients. Hence we define

Definition 3.2. For patient p, $age(p) = exit(p) - birth\ year(p) + 1$ with *exit* as in definition 3.1.

Since for some events the number of occurrences per patient is much correlated with the time spent in the system, some values must be featured as per year to achieve a comparable measure of frequency.

Definition 3.3. For patient p, $exposure(p) = exit(p) - entry(p) + 1$
where $entry(p) = \text{year}\,(entry\ date(p))$.

3.2.1. ICPC-coded levels

The episode list dates back the furthest of all tables, i.e. 1994 (the others were initialized in 2007), it is thus an important overview of childhood affairs for all the patients. However, according to the medical experts, it cannot be used as a trusted source of the number of such episodes or visits. This is because the rules for registering them vary among doctors and practices, as mentioned in 2.1 and 2.3. Some of the episodes reflect the events from the *consultations*, although not all of them, as they do not overlap fully. Therefore, to retain all available medical history, avoid cross-correlation with the number of visits and minimize GP-specific bias, the ICPC codes for episodes are included as one-hot encoded binary features. According to [73], the number of visits to a GP might be a good predictor for mental health problems. This information is exported from *consultations* as the number of unique dates a patient visited their doctor divided by exposure to have them counted per year.

The records that are not ICPC-coded contain binary responses for descriptive symptoms text mined from the GP's notes. Such features include *stress, behavioural problems, ADHD, autism* and *sleeping disorders*. The means of including these observations is a more intricate choice, as both chronic and acute symptoms can be found interweaving in alphabetical order. One option would be to count the visits with each syndrome recorded and divide it by exposure to express their yearly prevalence. Nonetheless, this operation introduces bias – for chronic diseases frequency of all appointments would be captured (as impeccably we assume these to be identified with every visit) and for non-recurring complaints it would be the frequency of the problem, diminished with the increase of exposure. The ways of reporting chronic and acute symptoms is already debatable and differs between GP-s. Consequently, together with the medical experts, the decision was made to incorporate this information as binary variables.

3.2.2. Ways of including measurements without values

As mentioned in section 2.3, it is unfeasible, at least for the purpose of this paper, to put the measurements into a framework that would enable to compare them between patients. Each such measurement, being either an assessment made by the GP or a laboratory result, is thus conveyed as the number of events per patient (without adjusting to exposure). The rationale is that in the Netherlands there is no obligation for cyclical blood tests so the counts are less dependent on the length of the available medical history. To extract even more information from the results, some additional features are created: number of having blood drawn for testing per year (which is the number

of unique dates with lab results in a patient's file, divided by exposure), minimal time between two tests of a kind and minimal time between two blood-sampling events. The hypothesis is that the amount of tests performed on an individual can be a good predictor, and it is based on the fact that the number of visits is in general a good proxy for poor health, as reported by [73].

The number of days between two tests may vary from 1 (if a test was repeated the next day) to infinity (if a test was done only once or has never been done) and infinity cannot be used in the models. Therefore there is a need of a function that converts it to some bounded interval. In function f below, the inverse provides the bounds and the logarithm distributes values more evenly.

Definition 3.4. Let $x \in [1, \infty)$ be the number of days between two tests. Transformation f is given by $f(x) = \frac{1}{\log(e+x-1)}$ which ensures $f(x) \in [0, 1]$ with $f(x) = 0$ for $x = \infty$ and $f(x) = 1$ for $x = 1$

The referrals, analogously to the measurements, are simply counted. In the Netherlands, a GP should refer a patient to a specialist once if there's a problem. This is ideally assumed but not always reflected by the data. The number of visits at the specialist is kept by the specialist, not the GP. Hence the information available to the general practice units (and the PIPPI study) about the consultations reflects on whether a patient has once visited a certain specialist or not and does not require dividing by exposure. One might righteously argue that the longer a patient has been in the system, the higher the chance of being referred. The claim that the number of each type of referrals can be expressed better as counts than counts per year is a simplifying assumption that provides a necessary trade-off.

3.2.3. Grouping medication

The last source to be described in this section provides input on prescribed medication in the form of ATC codes. Compared to the other features, this is the least biased by the GP's routines. The codes are constructed according to the function, type and composition of drugs, e.g. `D07AB02` (*hydrocortisone butyrate*) is listed as `D` for *dermatologicals*, `D07` for *corticosteroids*, `D07A` for *plain corticosteroids*, `D07AB` for *moderately potent corticosteroids* and the suffix `02` means it is the second drug in the last category. Especially with a sparse matrix, it is relevant to help the machine treat similar features as such [90] and there is a potential to group the medication according to the knowledge of the experts who came up with this framework. The segregation could be done by the machine itself, with methods such as Principal Component Analysis (PCA), it is however likely that the outcome would not be satisfactory considering the large number of zeroed dimensions.

The reason why grouping is important is the following. Let `A00AA01` and `A00AA02` be similar medicine that is not too common. A doctor would prescribe either `A00AA01` or `A00AA02` as there is no sense to take both at the same time. In the training set there are patients who were given `A00AA01` and the algorithm decides it is a fairly good indicator

of the outcome. However, the computer sees A00AA01 and A00AA02 being as far away as A00AA01 and B15CC05 or any other code, so a patient who takes A00AA02 will not be given any credit for using a similar medication while they should. Including a column indicating that a patient was given A00AA** (where * stands for any digit) would shift some weight allocated by the algorithm from A00AA01 to A00AA** and thus a person who was prescribed A00AA02 would be assigned higher score. Another advantage is that if there is an equal number of people taking A00AA01 and A00AA02 in the training set, the algorithm is more likely to put more weight on A00AA** category instead of distributing it between A00AA01 and A00AA02.

There might exist, however, a danger of grouping antagonistic medicine together. For example, in a situation where depression is correlated with amount of daylight [59], and so prescription of vitamin D could be an indicator [66], some high-level group could contain both vitamin D and a substance driving its levels down. Bringing these drugs together might increase false positive rate. In this study one category was carefully chosen with the specialists, grouping medication with the same first character and the following two digits (e.g. *hydrocortisone* D07AB02 and *betamethasone* D07CC01 fall into D07****) and reflecting the therapeutic subgroup (as called in table 2.2). These groups are referred to with a prefix *atc3*.

3.3. Pattern recognition

Using patterns in multivariate electronic health records serves to represent the temporal aspect of the data. The patterns are first defined, then mined, and finally added as features for the classification task. Each of these steps requires theorization ad development of a new approach tailored to accommodate the limitations of the data. The definition process must produce a set of consistent rules and a language that can represent the time dimension. Some additional assumptions, based on a supervised environment, are then imposed to reduce the number of patterns to be created. Although this is an uncommon step [108], since usually pattern mining is an instance of unsupervised learning, cutting down the number of features became an integral part of feature engineering when the improvements in data storage and collection over the past years have caused the necessity of adopting and processing huge datasets [4]. Preliminary selection of patterns or pattern elements based on their frequency and expected predictiveness helps to decrease the computational power needed and shorten the execution time. This procedure is repeated twice, first before the mining step, then to decide which patterns to include in the models. The patterns are passed to the models as binary features, representing whether a pattern was found in a patient's EMR or not.

3.3.1. Framework

The definitions introduced further in this chapter are inspired by [1], [4], and [83], except 3.5, 3.8, 3.10 and 3.13, which are custom-made and constitute the core of the methodology presented in this paper.

Let $P = \{1, ..., n\}$ be a set of index representations of patients identifying them explicitly and let $D = \{(x_i, y_i) : i \in P\}$ be a dataset such that $x_i \in \mathcal{X}$ denotes all available EMR for the i-th patient and $y_i \in \mathcal{Y} = \{0, 1\}$ is a class label corresponding to their medical condition. We call (x_i, y_i) an *instance* of D. It is worth noticing that (x_i, y_i) describes the i-th patient fully, hence the words *patient* and *instance* can be used interchangeably. The objective is to learn a function $f : \mathcal{X} \to \mathcal{Y}$ to predict the label with the highest accuracy according to the evaluation method. A map $\phi : \mathcal{X} \to \mathcal{X}'$, where $\mathcal{X}'$ is a linear space, is used to turn multirow EMR into values of an ordered set of features, with $x'_i = \phi(x_i)$ being a numerical vector of fixed length (corresponding to the number of features). The idea is to simplify and flatten the EMR records while preserving as much information as possible. The map ϕ does not have a closed form – it is the composition of all feature engineering transformations described in section 3.2 and the patterns. Let $X = \{x'_i : i \in P\} \subset \mathcal{X}'$ be a set of all images of D_1, which is a projection of D on its first coefficient and let $y = \{y_i : i \in P\}$ remain the respective target response. A regression table $X|y$ is then defined as a horizontally ordered matrix of $\{x'_i \in X\}$ and a corresponding vertical vector y. After separating P into P_{train} and P_{test}, models are fitted with $X_{train}|y_{train}$ and evaluated on $X_{test}|y_{test}$.

Some possible dimensions of $\mathcal{X}'$ were already defined in section 3.2 and introduction. Patterns, however, need a slightly different framework.

Definition 3.5. An event E is defined as an instance of an ICPC, ATC usage, referral to a specialist, a test being performed or a test result change (increase/decrease). For every ATC only a corresponding *atc3* group constitutes to an event. Let Σ be a finite set of all permitted events.

Definition 3.6. Let $E_i[t_{start}, t_{end}]$ be an event $E \in \Sigma$ registered for the i-th patient on date t_{start} and lasted until t_{end}. $t_{end} - t_{start} + 1$ indicates the duration of $E_i[t_{start}, t_{end}]$ in days.

The length of events is available in [57], a study of morbidity of ICPC-coded diseases. Five duration categories are defined: acute, lasting 4, 8 or 16 weeks, long-term (1 year) and chronic (no complaints-free period). Examples of such classification are presented in table 3.2. t_{end} is determined based on this research by adding 28, 56 and 112 days to t_{start} for the respective acute illnesses and 360 days for the long-lasting ones, to be consistent with the 180-day time window within which the records were neglected for target instances.

Being diagnosed with any chronic disease is a big and disturbing event in a child's life, and can greatly affect their mental health [11]. There is thus no need to consider subsequent records repeating this diagnosis, as they do not inflict any further stress or change the condition of the patient. In addition, it would be wrong to create patterns implying that an acute symptom preceded a chronic one, if the chronic issue was diagnosed earlier (and thus mishandle the order of events) which would happen if we allowed for duplicating episodes of chronic diseases. For these reasons only the date of identification is taken into account, or, if the disease was diagnosed before the patient registered, it is the date of the first record of this disease in the system. This date is used

as both starting and ending time of the event. For other variables (prescriptions, tests, referrals), $t_{start} = t_{end}$ is assumed to be the date of the visit when they were recorded.

One might argue if the durations provided by the study are accurate, as the four intervals are coarse-grained. It causes the ectopic pregnancy to be exemplified in table 3.2 as an instance of a 1-year affliction, yet a pregnancy cannot last so long [20]. However, in pursuance of mental health prediction it is less important to capture for how long precisely a person was bothered by direct symptoms of an illness than to determine for how long it was affecting them physically or mentally, for which the bounds are rather tenuous. An acute disease cured completely in a short time might have left a patient weak and exhausted, which would influence their quality of life and cause depression.

Table 3.2.: Examples of ICPC codes by morbidity

duration	code	description
4 weeks	B02	Enlarged lymph node(s)
	D82	Teeth / gum disease
	S09	Local infection of of finger(s) / toe(s)
8 weeks	H70	Otitis externa (inflammation of the ear canal)
	R05	Cough
	X85	Cervicitis or other cervix disease
16 weeks	A81	Multiple trauma / internal injuries
	S70	Post-herpetic neuralgia
	Z12	Abuse / sexual abuse by partner
1 year	D12	Constipation
	N19	Speech / phonation disorder
	W80	Ectopic pregnancy
Chronic	A28	Disability / handicap
	K74	Angina pectoris (coronary artery disease)
	L95	Ostheoporosis

Source: Calculation of morbidity figures based on NIVEL Care Registrations [57]

The interval-based representation can be simplified by duplicating E_i and linking it with dates between t_{start} and t_{end}. We get:

$$E_i[t_{start}, t_{end}] \mapsto \{E_{ij} : j \in \{t_{start}, t_{start} + 1, \ldots, t_{end}\}\}$$

for $t_{start} < t_{end}$ and

$$E_i[t_{start}, t_{end}] \mapsto E_{it_{start}}$$

for $t_{start} = t_{end}$.

Let $\{E_{ij}\}_j$ be a series of time-ordered events for a fixed i. The aim is to create temporal patterns that preserve this arrangement. Although for each $i \in P$ events $\{E_{ij}\}_j$ can be ordered by j, the order is never explicit. This is due to the fact that multiple events can be raised and registered during one visit (hence with the same date). As pointed out in section 2.3, the true order for such events is unknown.

Definition 3.7. An event sequence is a series of events where the events are ordered according to their start times:

$$\mathcal{E}_i = (E_{ij_1}, E_{ij_2}, \ldots, E_{ij_m}) \text{ s.t. } j_l \leq j_{l+1} \quad \forall\, l \in \{1, \ldots, m-1\}$$

Definition 3.7 provides a setup to think of a sequence as an ordered list of events where each event occurred no sooner than the preceding one. This tractable rationale is used to construct patterns by induction in definition 3.10. However, to better understand the mining algorithm described further in section 3.3.5, it helps to visualize sequences as ordered permutations of sets, as the algorithm operates on vectors rather than single events. The framework is presented below in definition 3.8. Here a set consists of events registered within one visit, characterized explicitly by its date, j. A permutation of any number of elements of the set is already a sequence (in line with definition 3.7). Such sequences derived from two sets with dates j_l and j_m respectively, such that $j_l < j_m$, can be concatenated preserving the order of the sets, first a permutation from the set indicated by j_l, then j_m.

Definition 3.8. Let $\sigma_1, \sigma_2, \ldots, \sigma_m$ be permutations and let $[k]$ denote $\{1, \ldots, k\}$. The following is an alternative definition of an event sequence:

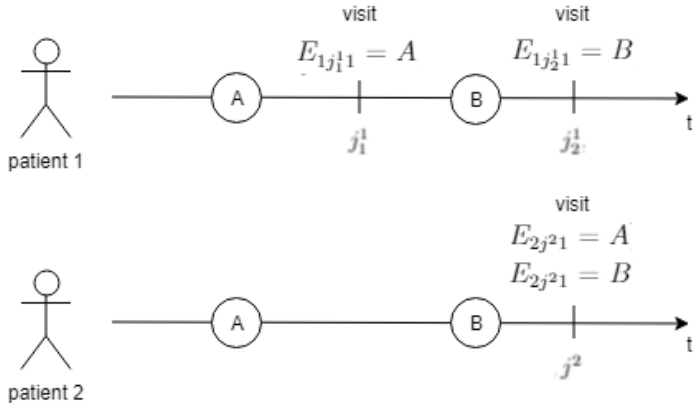
$$\mathcal{E}_i = \left(\sigma_1\left(\{E_{ij_1k} : k \in [k_{j_1}]\}\right), \sigma_2\left(\{E_{ij_2k} : k \in [k_{j_2}]\}\right), \ldots, \sigma_m\left(\{E_{ij_mk} : k \in [k_{j_m}]\}\right)\right)$$
$$\text{s.t. } j_l < j_{l+1} \quad \forall\, l \in \{1, \ldots, m-1\}$$

The i-th instance can be represented by an event sequence $\mathcal{E}_i$ and represented uniquely by a set $\{\mathcal{E}_i(\sigma) : \sigma = (\sigma_1, \sigma_2, \ldots, \sigma_m) \subset (S_{k_{j_1}}, S_{k_{j_2}}, \ldots, S_{k_{j_m}})\}$ where S_k is a group of permutations of k elements. Undoubtedly, a comprehensive framework for patterns should be independent of the choice of σ-s.

Definition 3.9. Events E_{ijk}, E_{ilm} are in relation R if $E_{ijk}\, R\, E_{ilm}$.

A relation between events that was chosen to be captured in this study can be best described by the linking word *before*, e.g. E_{ijk} *before* E_{ilm} which means that $j \leq l$. Notice how it is allowed for j and l to be equal (meaning they can be the same date). Since this is the only relation considered, the word *before* and the symbol R can be omitted for shorter notation (E_{ijk}, E_{ilm}). This approach serves mainly to capture situations such as depicted in figure 3.1: patients (1) and (2) observed events A and B in this order, with patient (1) reporting them to their GP consecutively, right after they happened, and (2) mentioning them both during one appointment. The EMR of the patients contain $\{E_{1j_1^1 1} = A, E_{1j_2^1 1} = B\}$ and $\{E_{2j^2 1} = A, E_{2j^2 2} = B\}$ where A, B are events and j_1^1, j_2^1, j^2 represent the dates of visits. Patterns (B, A) and (A, B) are created in the second instance and this allows to pick up the similarity with a single pattern (A, B) mined for instance (1). Definition 3.10 expresses this logic in a formal way and figure 3.2 graphically illustrates how patterns (`icpc_D10,icpc_n89,icpc_R78,atc3_J01`) and (`slaapstoornis,ver_longarts,icpc_R96`) can be created from an exemplary timeline of the i-th patient.

Figure 3.1.: Timelines with the same events and different distribution of visits



Definition 3.10. Let E_{ijk} denote the k-th event registered for the i-th instance on the j-th date.

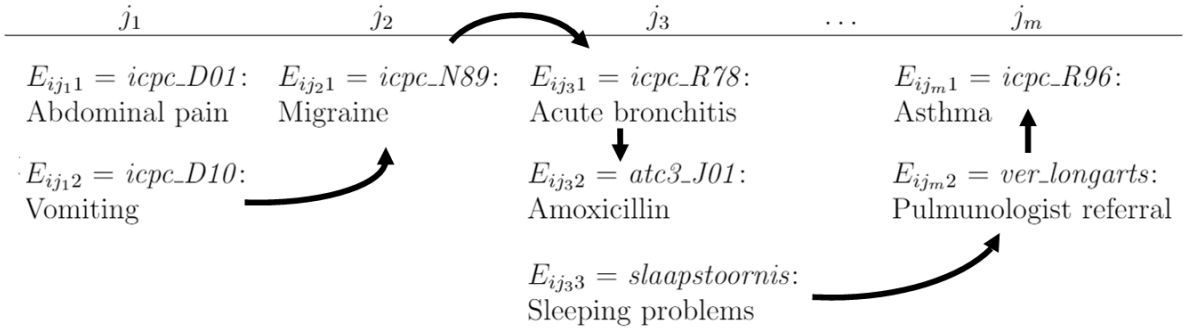
i) The size of a pattern $\mathcal{P}$ is the number of events it contains. If the size of $\mathcal{P}$ is n, we call $\mathcal{P}$ an n-pattern. The space of all temporal patterns of arbitrary size is denoted by $\mathcal{TP}$ and all patterns mined from the i-th instance are referred to by $\mathcal{T}_i$.

ii) A 1-pattern is defined by E if $\exists\, j, k$ s.t. E_{ijk} exists for some $i \in P_{train}$. We say that E_{ijk} exists if $\exists\, e \in \Sigma$ s.t. $E_{ijk} = e$.

iii) A 2-pattern is constructed from 1-patterns as (E_{ijk}, E_{ilm}) where $l \geq j$ and $E_{ijk} \neq E_{ilm}$. We say E_{ijk} happened *before* E_{ilm}.

iv) An n-pattern is constructed by induction. Let $\mathcal{P} = (E_{ijk}, \ldots, E_{ilm})$ be an $n-1$ pattern. If $\exists\,, q$ s.t. E_{ipq} exists and $p \geq l$ and $E_{ipq} \neq E_{ijk} \wedge \ldots \wedge E_{ipq} \neq E_{ilm}$ then $\mathcal{P}' = (E_{ijk}, \ldots, E_{ilm}, E_{ipq})$ is an n-pattern. We call $\mathcal{P}$ a subpattern of $\mathcal{P}'$ and note $\mathcal{P} \subset \mathcal{P}'$.

Figure 3.2.: Examples of a 3-pattern and a 4-pattern on a timeline



3.3.2. Controlling the number of patterns

Let $\mathfrak{n}$ be the power of Σ. Then, there are $\mathfrak{n}$ 1-patterns, $\mathfrak{n}^2 - \mathfrak{n}$ 2-patterns, $\frac{\mathfrak{n}!}{(\mathfrak{n}-k)!} = \binom{\mathfrak{n}}{k} \cdot k!$ k-patterns that can be created using events in Σ for $k \leq \mathfrak{n}$. It can easily be shown that $\binom{\mathfrak{n}}{k} \cdot k! \geq \mathfrak{n} \cdot k!$, therefore without any restrictions the number of patterns witnesses factorial growth with the increase of the length of the patterns and cardinality of Σ. To make it worse, for EMR data $\mathfrak{n}$ is a large number, as portrayed in table 4.3. It is computationally ineffective to mine all the patterns that may exist.

Definition 3.10 (*ii*) contains an important condition that helps reducing the number of mined patterns: a pattern must exist in the medical history of at least one of the patients in a given set. This set is chosen to be the training set in order not to contaminate the models with information from the test set. It also helps to approximate the setup where the potential classification tool would be used.

Definition 3.11. The support of pattern $\mathcal{P}$ in P_{train} is the number of instances in $D|_{P_{train}} = \{(x_i, y_i) \in D : i \in P_{train}\}$ that contain $\mathcal{P}$:

$$supp(\mathcal{P}, D|_{P_{train}}) = |\{i \in P_{train} : \mathcal{P} \in \mathcal{T}_i\}|$$

It is clear that

$$\forall\, \mathcal{P}, \mathcal{P}' \in \mathcal{TP} \quad \mathcal{P} \subset \mathcal{P}' \Rightarrow supp(\mathcal{P}, D|_{P_{train}}) \geq supp(\mathcal{P}', D|_{P_{train}})$$

Definition 3.12. A temporal pattern $\mathcal{P}$ is called *incoherent* if there does not exist any valid event sequence $\mathcal{E}_i$ for $i \in P_{train}$ s.t. $\mathcal{P} \in \mathcal{T}_i$. In other words, $\mathcal{P}$ is incoherent $\Leftrightarrow supp(\mathcal{P}, D|_{P_{train}}) = 0$.

It is irrelevant to mine incoherent patterns, as they would not convey any information for the models. Similarly, it is unnecessary to mine patterns that are very rare or carry ambiguous information. We say that a pattern is target-specific if it accounts for a high risk of being classified as target and non-target-specific if it increases the chance of being recognized as non-target. Both target and non-target-specific patterns are useful in a classification task. On the other hand, non-specific patterns, i.e. patterns that are equally common for both labels of the response variable, are unlikely to be assigned much weight by an AI classifier and can thus be removed from the models without a loss of substantial information. The following definitions describe a method of measuring specificity of patterns.

Definition 3.13. Let $\mathcal{P}$ be a coherent pattern and let $y_i \in \mathcal{Y}$ for some $i \in P_{train}$. Define $\mathcal{P}_i = y_i$ if $\mathcal{P} \in \mathcal{T}_i$. For i s.t. $\mathcal{P} \notin \mathcal{T}_i$ $\mathcal{P}_i$ is not defined. Then $\mathcal{P}_i$ is a random variable with values in $\mathcal{Y}$.

Definition 3.14. The confidence of $\mathcal{P} = \tilde{y}$ is the proportion of instances from class $\tilde{y}$ in all instances covered by $\mathcal{P}$ in the training set where $\tilde{y}$ is a value of the label.

$$conf(\mathcal{P} = \tilde{y}) = \frac{supp(\mathcal{P}, D|_{P_{train}, y=\tilde{y}})}{supp(\mathcal{P}, D|_{P_{train}})}$$

which is no different from a sample probability $\mathbb{P}_{sample}(\mathcal{P}_i = \tilde{y})$

It is straightforward that the higher the confidence of $\mathcal{P} = 1$, the more target-specific $\mathcal{P}$ is and analogously the higher the confidence of $\mathcal{P} = 0$, the more non-target-specific it is. It is easy to notice that $conf(\mathcal{P} = 0) = 1 - conf(\mathcal{P} = 1)$. This sample probability is a good approximation of what is precisely that we want to know for each temporal pattern – the true probability that a patient is ill on condition that we observed this pattern in their medical history. Moreover, the following holds:

Theorem 3.15. *$conf(\mathcal{P} = \tilde{y})$ is a maximum likelihood estimator (MLE) of the conditional probability $\mathbb{P}_{sample}(\tilde{y}|\mathcal{P})$*

Proof. Let $\mathcal{P}$ be a pattern and let $P^{\mathcal{P}}_{train} = \{i \in P_{train} : \mathcal{P} \in \mathcal{T}_i\}$. Pursuant to definition 3.13,

$$\mathcal{P}_i = 1 \quad \Leftrightarrow \quad \mathcal{P} \in \mathcal{T}_i \wedge y_i = 1 \quad \Leftrightarrow \quad i \in P^{\mathcal{P}}_{train} \wedge y_i = 1$$

As it is the conditional probability to be estimated, assume further $i \in P^{\mathcal{P}}_{train}$ holds (this is possible as $\mathcal{P}$ is coherent). Denote $p = \mathbb{P}_{P^{\mathcal{P}}_{train}}(\mathcal{P}_i = 1)$ and so $1 - p = \mathbb{P}_{P^{\mathcal{P}}_{train}}(\mathcal{P}_i = 0)$. $\mathcal{P}_i$ is therefore a random variable with Bernoulli distribution, which is a special case of binomial distribution: $\mathcal{P}_i \sim Bin(1, p)$. Accordingly,

$$f_i(x) = \mathbb{P}_{P^{\mathcal{P}}_{train}}(\mathcal{P}_i = x) = p^x(1-p)^{1-x}$$

where f_i stands for the probability mass function (pmf) of $\mathcal{P}_i$. $\mathcal{P}_i$ are i.i.d. and thus we can calculate:

$$\mathcal{L}(p, x) = \prod_{i \in P^{\mathcal{P}}_{train}} \mathbb{P}_{P^{\mathcal{P}}_{train}}(\mathcal{P}_i = x_i) = p^{\sum_{i \in P^{\mathcal{P}}_{train}} x_i} (1-p)^{|P^{\mathcal{P}}_{train}| - \sum_{i \in P^{\mathcal{P}}_{train}} x_i}$$

$$\ell(p, x) = \log(\mathcal{L}(p, x)) = \log(p) \left(\sum_{i \in P^{\mathcal{P}}_{train}} x_i \right) + \log(1-p) \left(|P^{\mathcal{P}}_{train}| - \sum_{i \in P^{\mathcal{P}}_{train}} x_i \right)$$

$$\frac{d}{dp}\ell(p, x) = \frac{1}{p} \sum_{i \in P^{\mathcal{P}}_{train}} x_i - \frac{1}{1-p} \left(|P^{\mathcal{P}}_{train}| - \sum_{i \in P^{\mathcal{P}}_{train}} x_i \right) = 0$$

Assuming $p, 1 - p > 0$ we arrive with MLE

$$\hat{p} = \frac{\sum_{i \in P^{\mathcal{P}}_{train}} x_i}{|P^{\mathcal{P}}_{train}|}$$

where the numerator is precisely $supp(\mathcal{P}, D|_{P_{train}, y=1})$ and the denominator is by definition the number of instances in the training set that contain $\mathcal{P}$ and so is $supp(\mathcal{P}, D|_{P_{train}})$. For $p = 1$:

$$\hat{p} = \frac{\sum_{i \in P^{\mathcal{P}}_{train}} x_i}{|P^{\mathcal{P}}_{train}|} = \frac{\sum_{i \in P^{\mathcal{P}}_{train}} 1}{|P^{\mathcal{P}}_{train}|} = \frac{|P^{\mathcal{P}}_{train}|}{|P^{\mathcal{P}}_{train}|} = 1$$

and for $p = 0$:

$$\hat{p} = \frac{\sum_{i \in P^{\mathcal{P}}_{train}} 0}{|P^{\mathcal{P}}_{train}|} = 0$$

□

Nevertheless, a very important observation must be made about the support of $\mathcal{P}$.

Remark. If $supp(\mathcal{P}, D|_{P_{train}, y=\tilde{y}}) = supp(\mathcal{P}, D|_{P_{train}}) = 1$ then $conf(\mathcal{P} = \tilde{y}) = 1$.

This shows that confidence is not a good measure of specificity for patterns that are rare in both classes of the response variable. A search of highly predictive patterns must promote patterns that are both specific and sufficiently common in classes they represent. If the selected patterns are not specific, they will not help to separate the instances, and if they are not common, they will only help to separate a few. In both cases the patterns will fail to noticeably improve the performance of Machine Learning models.

3.3.3. Comparison with Batal's approach

As mentioned in section 3.3.1, the methods used in this research build up on the theory described in [4], which itself is based on [1]. The latter lists 7 different relations (13 if we reverse the order of the linked events) which have widely been used [28, 32, 55, 64, 106]. However, with the EMR data, not all of them can be applied because the exact starting and ending times of the events are unknown. The trend is therefore to limit the number of possible links and Batal [4] claims that only two of them are relevant: *before* and *co-occures with*. In this paper, although only the first one is used explicitly, both relations can be captured. This is done by allowing events recorded within one visit to be connected in any order. All patients who reported to have had an event A co-occurring with B will have indication of patterns (A, B) and (B, A).

A lack of a window within which a patient is observed stands for another aspect distinguishing this methodology from Batal's (and [28]). If the EMR for the i-th instance consist of the medical history from t_0^i until t_1^i, it might be desirable to look only at those from a fixed interval $[\max(t_1^i - w, t_0^i), t_1^i]$ where w denotes the length of the window. Such an approach resolves most of the problems with feature engineering of whether to include or exclude exposure, making different instances more comparable. Furthermore, as indicated in [94], recent measurements of the clinical variables are more predictive. Most diseases, especially those for which decision support is being sought, evolve or give first symptoms in a short, defined time. This is not true for mental health disorders though. They can emerge rapidly or be caused by a distant childhood experience, such as death in the family [12, 35, 110], or a recurring one, e.g. shortage of daylight [59]. All available data is therefore recommended to be taken into account when dealing with mental issues.

The following defines a *rule* – a concept used by Batal to symbolize that a pattern is likely to imply a certain outcome. A kindred notion is grasped by definition 3.13.

Definition 3.16. A rule is defined to be of the form $\mathcal{P} \Rightarrow y$, where $\mathcal{P}$ is a temporal pattern and y is a specific value of the target class variable. We say that rule $\mathcal{P} \Rightarrow y$ is a subrule of rule $\mathcal{P}' \Rightarrow y'$ if $\mathcal{P} \subset \mathcal{P}'$ and $y = y'$.

Batal then describes confidence of $\mathcal{P} \Rightarrow y$ as in 3.14. This, however, is not precise. The notation suggests logical implication, so $\mathcal{P} \Rightarrow y$ would be true for instances $\{i : \mathcal{P} \in \mathcal{T}_i \wedge y_i = y\}$, but also for all patients who do not witness $\mathcal{P}$ and belong to either of the classes, y or its opposite. Unless it is assumed that we limit the set to only those instances for which $\mathcal{P} \in \mathcal{T}_i$ holds, it is not straightforward to arrive with the sample conditional probability that is used to score the patterns. Such assumption is used implicitly in Batal's definition of confidence but not in definition 3.16, one may thus find this documentation slightly misleading. The idea to substitute the rule with a more complex idea, extending it with random variables $\mathcal{P}_i$ and theorem 3.15, is an attempt to make the theory more transparent.

In addition to the above fixes, this paper has three main contributions to the approach: an efficient enrichment technique, an improved scoring method based on Wilson confidence interval and a framework to mine patterns in a setting where events may overlap, together with a new mining algorithm. The enrichment, as described in section 3.3.2, allows for a conversion from events as intervals to events as points in time. Not only does it simplify the concept (for only 3 pieces of information are needed to define an event instead of 4), but also creates possibility of vectorizing the mining algorithm. Moreover, it can be shown that the duplicating process can be limited to only the dates already existing in EMR data (see theorem 3.18 in section 3.3.5).

The improved scoring method tackles the problem of using confidence as a measure of specificity depicted by a remark at the end of the previous section. Batal deals with it by imposing a fixed threshold of instances that a pattern must cover to be considered. The threshold, denoted σ_y, is called *minimal support* and is chosen locally per class, i.e. the values differ for target and non-target. In the process of mining, first all candidates for patterns are created, then the ones with support lower than σ_y are removed. The threshold is chosen arbitrarily without any further assumptions and most likely dependent on the magnitude of the sample. This step is thus intractable and irreproducible, and for these reasons a new and slightly more advanced approach is proposed and applied in this study.

3.3.4. Wilson score interval

Wilson score is an example of a binomial proportion confidence interval and was first described in [103]. As the name suggests, it is used with binomial distributions. Such assumption can be made for the target and diseases in general for the reasons given by [23]. Firstly, a binomial event has exactly two possible outcomes and if the probability of one is known, the probability of the other is the difference of that probability from 1. The state of a single patient can indeed be described as either healthy or not healthy regarding a particular issue or a set of issues. Trials, i.e. testing different patients one by one for the target illness, are rather statistically independent. A counterexample to

this statement would be if the test was for a disease that has local outbursts, e.g. lice in a kindergarten (then if one child has it, the chance is high that the others have it too), yet for mental health the lack of correlation can be assumed, as it is not contagious. A *success* of a trial is defined by diagnosing the target illness. Altering the number of successes within a fixed number of trials, which is the sample size, is an instance of binomial distribution.

One much celebrated property of the binomial distribution $Bin(n, p)$ is that it asymptotically converges to a normal distribution $\mathcal{N}\left(np, \sqrt{\frac{1}{n}p(1-p)}\right)$ by Central Limit Theorem. The Wilson score interval is an extension of the normal approximation and improves precision of coverage probability of the interval. The nominal coverage probability is simply the confidence level set by the definition of the interval, e.g. it is 0.99 for error rate $\alpha = 0.01$. The actual coverage probability, on the other hand, is the true probability that the interval contains the value it was aimed for. A discrepancy between the actual coverage probability and the nominal coverage probability occurs mainly when approximating a discrete distribution with a continuous one which is exactly the case when estimating binomial experiment with a normal random variable.

The resemblance to normal distribution is employed to lower confidence from definition 3.14 by including the error of calculating the conditional probability from the sample. The following definition gives both bounds for the interval but only the lower bound is used.

Definition 3.17. Wilson score interval is given by

$$\frac{1}{1+\frac{z^2}{n}}\left(\hat{p}+\frac{z^2}{2n} \pm z\sqrt{\frac{1}{n}\hat{p}(1-\hat{p})+\frac{z^2}{4n^2}}\right)$$

where $z = 1-\frac{\alpha}{2}$ is a quantile of two-tailed standard normal distribution and α is a chosen error rate.

The lower the α, the more permissive the interval, hence decreasing the error rate scales down the lower bound. In this research α is chosen to be 0.01 as recommended in [74]. Examples of how this transformation lowers the initial score can be found in table 3.3.

Table 3.3.: Examples of Wilson score lower bounds

$supp(\mathcal{P}, D\|_{P_{train}, y=1})$	$supp(\mathcal{P}, D\|_{P_{train}})$	$conf(\mathcal{P}=1)$	Wilson
78	120	0.6500	0.5325
6	6	1.0000	0.4741
5	5	1.0000	0.4289
7	8	0.8750	0.4242
66	123	0.5366	0.4217
4	4	1.0000	0.3754
10	15	0.6667	0.3491
2	3	0.6667	0.1439

Source: own calculation

This modified scoring serves to choose the best $(n-1)$-subpatterns for creating n-patterns. Patterns with the highest scores are the most target-specific and with the lowest (usually 0) – the most non-target-specific. If two patterns have the same score but different support, the one with bigger support is considered better. After creating n-patterns, the scoring function determines features to be used in the models.

3.3.5. Mining algorithm

Describing what the patterns look like does not yet define how to mine them. The biggest challenge is to reduce the number of the patterns while retaining their usefulness. Batal provides that limitation by proving that there are at most $n+1$ coherent candidates that result from extending an n-pattern with a new event. Nevertheless, a completely new mining algorithm is necessary for this research, as his proof relies on the assumption that certain groups of events cannot overlap. It is clear that all kinds of events can overlap when dealing with data sourced from GP-s, as they can only be reported during appointments.

In the first step of the algorithm the events are collected directly from EMR by selecting the patient's identification number, name of the event and the date of the visit during which the event was reported. Then some events are duplicated in line with section 3.3.1 to account for their duration. For the sake of computation, only the dates that already exist in the i-th patient's medical history (the dates of visits) are used to enrich the event sequence with the lengths of ICPC-coded events.

Theorem 3.18. *Duplicating events with dates within their duration time that otherwise do not appear in the patient's event sequence does not change the patterns mined from that sequence.*

Proof. Let i be a fixed instance in P_{train} and $E_i[t_{start}, t_{end}] = E$ be the event of interest with $t_{start} \neq t_{end}$. Let $J = \{j \in x_i|_{start\ date} : x_i \in D_1\}$. J is then a set of all dates of visits in the EMR of i. Let j_0 be s.t. $j_0 \in (t_{start}, t_{end}]$, $j_0 \notin J$ and $\exists\, j_{-1}, j_1 \in J$ s.t. $j_{-1} < j_0 < j_1$ and $j_{-1}, j_1 \in [t_{start}, t_{end}]$, and $\nexists\, j'_{-1}, j'_1 \in J$ s.t. $j_{-1} < j'_{-1} < j_0 < j'_1 < j_1$. In other words, j_0 is a date between two consecutive dates in the i-th patient's EMR. Assume $\mathcal{E}_i$ is enriched with E on all dates in $\{t_{start}, t_{start}+1, \ldots, t_{end}\}$, including j_{-1}, j_0, j_1. The following elements of $\mathcal{E}_i$ are created (note, however, that j_{-1} may be equal to t_{start} and $k_{t_{end}}$ is 0 if $t_{end} \notin J$):

$$\{E_{i,t_{start},k_{t_{start}}+1}, \ldots, E_{i,j_{-1},k_{j_{-1}}+1}, \ldots, E_{i,j_0,1}, \ldots, E_{i,j_1,k_{j_1}+1}, \ldots, E_{i,t_{end},k_{t_{end}}+1}\}$$

We consider 3 cases in which $E_{i,j_0,1}$ could be used to form a pattern:

i) E is at the end of a pattern (creates *superpatterns* from existing patterns). Such patterns already exist with $E_{i,j_{-1},k_{j_{-1}}+1}$ at the end, which is also equal to E. The only way $E_{i,j_0,1}$ could make a change is if it would be added after $E_{i,j_{-1},k_{j_{-1}}+1}$ but this is not allowed according to definition 3.10, as $E_{i,j_{-1},k_{j_{-1}}+1}$=E=$E_{i,j_0,1}$.

ii) E is in the beginning of a pattern. This is only allowed if no event E is contained in that pattern after $E_{i,j_0,1}$. Let $\mathcal{P}$ be such a pattern where $E{=}E_{i,j_0,1}$. Then patterns created from $\mathcal{P}$ by replacing $E_{i,j_0,1}$ with $E_{i,j_{-1},k_{j_{-1}}+1}$ or $E_{i,j_1,k_{j_1}+1}$ are the same as $\mathcal{P}$ and still preserve the order.

iii) E is in the middle of a pattern. A similar argument holds as above – if $E_{i,j_0,1}$ builds any pattern, then neither $E_{i,j_{-1},k_{j_{-1}}+1}$ nor $E_{i,j_1,k_{j_1}+1}$ appear respectively before or after $E_{i,j_0,1}$ and they can be used to create the same pattern by replacing $E_{i,j_0,1}$.

It has been proved above that for each pattern mined using $E_{i,j_0,1}$ there exists an identical pattern created with $\{E_{ijk}\}_{j,k}$ s.t. $j \in J$. Hence duplicating E as $E_{i,j_0,1}$ when $j_0 \notin J$ does not change the set of patterns mined from instance i. □

The measurements from *tests* are divided according to the type of their outcome: numerical or categorical. Numerical events are sorted by patient, name of the test and the date of having it performed. Differences between consecutive values are computed to determine whether the results were increasing or decreasing in time. Test names having positive and negative differences are marked with prefixes *inc_* and *dec_* respectively.

Definition 3.19. Let M_{ij} be a numerical value of the outcome of a test M performed on the i-th patient on date j and let M_{il} analogously be the result of the same type of a test done on date l. Assume $\nexists M_{ik}$ s.t. $j < k < l$. In other words, M_{ij} and M_{il} are chronologically consecutive for patient i and test M. Then a difference $M_{il} - M_{ij}$ is calculated and for some $\delta \geq 0$:

i) an event $E_{il*} = inc\_M$ is created if $M_{il} - M_{ij} > \delta$

ii) an event $E_{il*} = dec\_M$ is created if $M_{il} - M_{ij} < -\delta$

iii) no event is created if $-\delta \leq M_{il} - M_{ij} \leq \delta$

where $*$ stands for an indexing number of the event created with date l, e.g. if E_{il*} is the first event defined with date l for the i-th patient, we put $* = 1$. In this study $\delta = 0$ is assumed.

As for the tests with textual response, it is difficult to separate the outcomes into positive and negative as mentioned in section 2.3, hence only the fact of having the test done, which is already a fair indicator of the GP's suspicion, is captured as an event.

The events are scored and the most target-specific and non-target-specific of them are chosen. The idea of restricting the number of subpatterns before mining longer patterns comes from [27]. For each patient all events are then sorted by date. For a fixed patient the second step involves creating two lists containing the name of the first (earliest) event recorded for that patient: one called *prev* for previous events and the other, *prev_date*, for events that were recorded on the same date. The date itself is stored as a 1-dimensional variable *date*. Additionally, two other lists are created, *prev_good* and *prev_date_good*. If the first event is considered *good*, i.e. selected as a pattern of high specificity, it is added to those lists as well.

The third step is iterating through each of the tuples (event,date) of that same patient in chronological order. If the date is the same as *date*, i.e. the events were raised during the same visit, and the event is *good*, then it matches on the right hand side with all the previous (*prev*) events and on the left with all the *prev_date*. Then it is added to the *good* lists. If it is not a specific pattern, then it matches analogously with *prev_good* and *prev_date_good*. In case the dates differ, the lists *prev_date* and *prev_date_good* start over and the event is joined only with either of the others. This step is depicted by a flowchart in figure 3.3 on page 30. Horizontal boxes contain transformations on variables, hexagons mark if clauses, vertical rectangles show how to concatenate vectorized lists to form 2-patterns and *len* is short for length of a vector.

An important feature of this algorithm is that it only involves data gathered from one patient at a time. It is therefore encouraged to run it in parallel. Another advantage is that the patterns are added as vectors (or a list of tuples), not one by one. The part described above is presented in a python code snippet A.1 in the appendix.

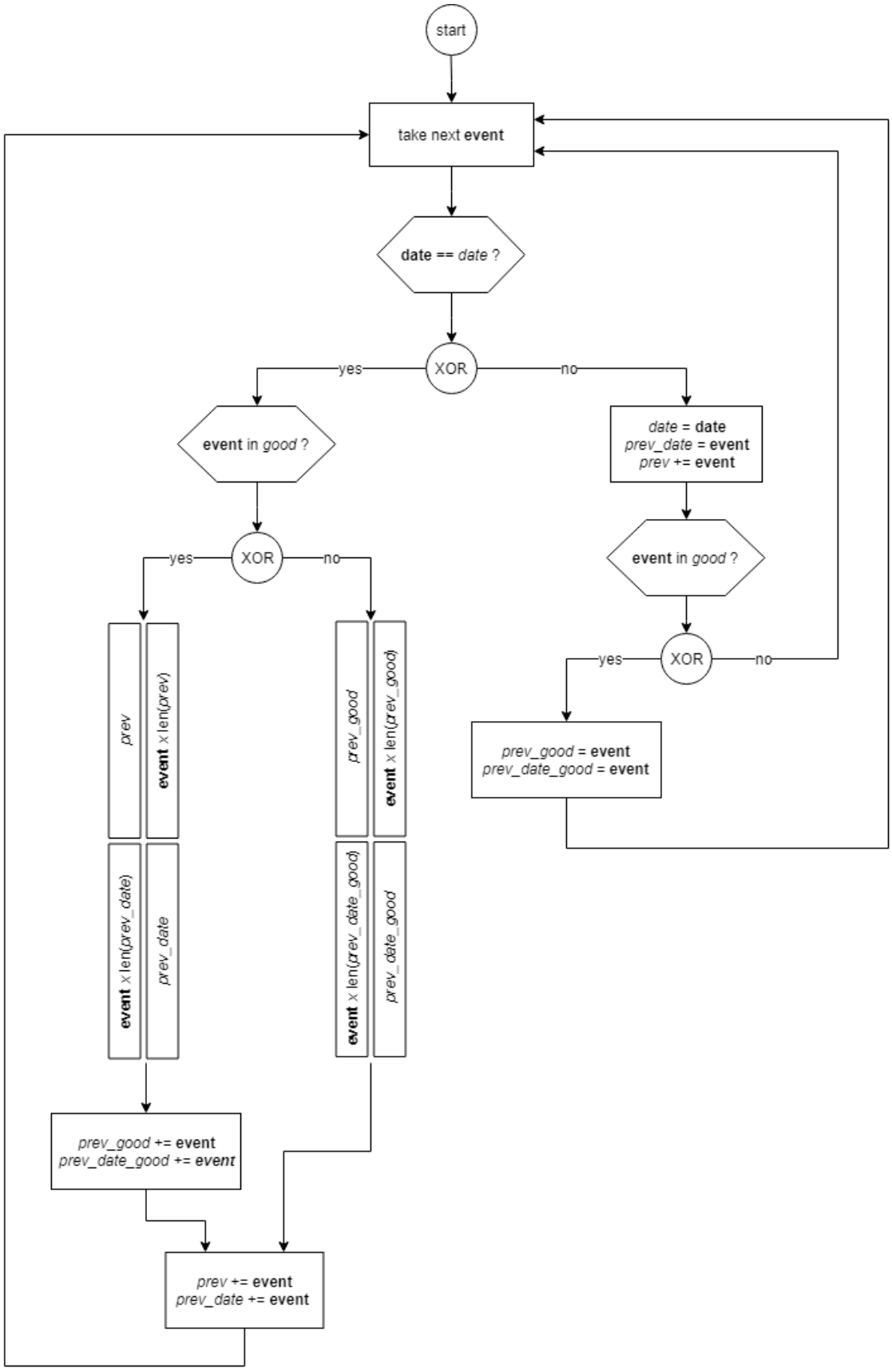
The date of the second event in a 2-pattern is saved alongside the pattern for the purposes of mining longer patterns. In that case, an event is added at the end to form a superpattern and the same rules apply as with creating 2-patterns. An algorithm for mining patterns longer than 2 is by far less complex. An example in python is attached in the appendix under A.2. Since the patterns contribute to the models as binary features, spawning identical superpatterns can be significantly reduced by dropping duplicates per instance before starting the algorithm. This step can be applied no sooner than the 2-patterns have been created. The pattern that is kept is always the earliest one, otherwise potential superpatterns would be lost.

3.4. Evaluation methods

To fully grasp the idea behind the choice of the methods used with the models, it is necessary to understand what constitutes to a good model and what measures are being validated. The PIPPI project has two major assumptions regarding this evaluation. For each acceptable model, *sensitivity* (ratio of true positives to all positives in a sample, also called True Positive Ratio, TPR) and *specificity* (ratio of true negatives to all negatives) should both be higher than 0.7, which means that 70% of healthy and 70% of unhealthy people should be correctly recognized [39].

Such models will be then compared using the value of AUC which is the area under ROC curve (Receiver Operating Characteristic). The curve is a function of thresholds, i.e. cut-off points that decide how continuous outcomes of a model should be separated into classes. The vertical and horizontal axes are TPR and FPR respectively (False Positive Ratio which is a ratio between false positives and all negatives) and it is desirable that the curve lies above the $FPR = TPR$ level, which would be a curve for a random classifier. AUC$\in [0, 1]$ is the euclidean area under this curve, with 0.5 being a mediocre score (no better than random) and 1 being a perfect score. An efficient algorithm is able to assign the highest probabilities to many target instances and the lowest to the majority of the non-target ones, and this concept is captured by the ROC.

Figure 3.3.: Iterative part of the mining algorithm



To compare AUC values of two models a z-statistic test can be conducted as described by DeLong [21]. The null hypothesis of a two-tailed test claims that the AUC values are equal and should be rejected if the p-value is lower than $\alpha = 0.05$.

3.4.1. Criticism of ROC

This section argues that using ROC curve and the AUC as the only evaluation method is flawed in mental health diagnostics. In this specific area, it is more acceptable to arrive with many false positives than a lot of false negatives. This is because depression and other issues are already heavily underrecognized and undercured, and the cost of mistakenly alerting a GP with a problem is minimal - it results in asking more questions or a visit to a psychologist. In another setting, the opposite might be appropriate: if Artificial Intelligence was used to judge in a court, it would be more socially acceptable for it to leave a criminal at large than to put an innocent person in jail. Although very different, both situations are treated equally by ROC and AUC.

Furthermore, ROC depends much on how balanced the dataset is. In this particular case, with 27% target instances, lowering the threshold to arrive with one more true positive and two or three false positives is almost as good as no change at all (if we started from the diagonal line $FPR = TPR$, we would move along it and still perform no better than random). So the trade-off between false positives and true positives is given by the proportion of negatives to positives in the sample and cannot be changed. This is another disadvantage of the ROC, since it often occurs in Machine Learning that the outcomes should be weighted (or penalized) differently.

Regression trees and random forests segregate patients into buckets (called leaves or final nodes) using a collection of classification rules that is the outcome of these methods. Each leaf is assigned probability that a mental issue is going to be developed so the values of possible responses are discrete and ordered, and their number is finite. Thus the graphs of ROC curves will always be piecewise linear, bounding smaller areas than a fully concave function would. This poses an inconvenience when comparing trees to other models using AUC.

3.4.2. Evaluating the impact of patterns

The PIPPI project does not impose any means of comparison between models that are trained with a different number of features. Such evaluation is required to assess the impact of the patterns. Although usually the more features in a model, the better its performance, the key is to reduce dimensionality with little loss of efficiency. This is done by feature engineering and choosing parameters that carry, but not repeat, the most substantial information. Typical measures of relative quality that involve the number of features and sample size are Akaike Information Criterion and Bayesian Information Criterion.

Definition 3.20. Let m be the number of parameters, n be the sample size and $\hat{\mathcal{L}}$ the maximized likelihood function. Then:

$$AIC = 2m - 2\log(\hat{\mathcal{L}})$$

$$BIC = m\log(n) - 2\log(\hat{\mathcal{L}})$$

As mentioned in section 3.3.4, binomial distribution is a fairly justified approximation of the target variable. In the proof of theorem 3.15 it was calculated that $\mathcal{L} = p^{\sum x}(1-p)^{n-\sum x}$ satisfied by MLE $\hat{p} = \frac{|P_{train,y=1}|}{|P_{train}|}$. For evaluation we could put $\sum x = |P_{test,y=1}|$ and $n = |P_{test}|$. However, such measure would only give information about how many positive instances were predicted in the test sample, with $\hat{p}$ constant across all models built on P_{train}, which is not sufficient to compare them. It is therefore practical to substitute likelihood with a loss function based on L_1 metric.

Definition 3.21. Let $y_{test} = \{y_i : i \in P_{test}\}$ be the true response vector and let $\hat{y}_{test} = \{\hat{y}_i : i \in P_{test}\}$ be a vector of responses predicted by an algorithm.

$$\mathcal{L} = \sum_{i \in P_{test}} |y_i - \hat{y}_i| = \sum_{i \in P_{test}} (y_i - \hat{y}_i)^2$$

As the aim is to minimize the loss and the indicators penalize the number of features and observations with positive values, the signs must be modified as follows [33]:

Definition 3.22. Let m be the number of parameters, n be the sample size and $\mathcal{L}$ the residual sum of squares. Then:

$$AIC = 2m + n\log\left(\frac{\mathcal{L}}{n}\right)$$

$$BIC = m\log(n) + n\log\left(\frac{\mathcal{L}}{n}\right)$$

AIC or BIC, defined by 3.21 and 3.22 together, can be employed to assess models trained on the same sample but varying in number of features, and BIC can additionally be used to compare models with different sample sizes, for instance between age groups.

3.5. Models

Six classifiers are chosen to process the data: logit, SVM, regression tree, random forest, deep neural network and XGBoost. Logit, i.e. logistic regression, is a special case of a generalized linear model with a linking function that converts probability to the logarithm of odds ratio. It returns a value on a $[0, 1]$ scale that can be used to classify the instances. The Support Vector Machine serves to find a separating hyperplane that

maximizes the Euclidean distance between classes represented in linear space. Similarly, neural networks aim to find a hyperplane to separate classes, with each layer of nodes training on aggregated output from the previous layer [27]. A regression tree outputs a set of rules based on values of chosen features. Each rule separates a node into child nodes in a decision tree manner. Final nodes contain instances assigned the same output value. Random forests stack many regression trees and for each instance they choose a class that was assigned to that instance the most. Gradient boosting is used to enhance performance in XGBoost – trees are ensembled one by one to correct the errors made by the existing trees [16].

The models are cross-validated with 3-folds and all but the first are tuned for AUC using GridSearch or RandomSearch, depending on the number of hyperparameters available and relevant to be tried. This section does not delve into the Machine Learning algorithms but explains how to apply them to healthcare data and the requirements of the PIPPI project.

An ideal scoring function for such task would penalize false positives and false negatives differently, preferably according to economical losses in case of the false positives (e.g. the cost of psychiatric consults) and financial valuation of the combined social and economical consequences of an untreated mental issue (e.g. a child struggling with depression or anxiety might take more days off in the future or work less effectively than their healthy peers [6]). Such estimation would, however, require a separate study, so at the moment it is unavailable. AUC was chosen as evaluation function, as maximizing it helps to achieve high specificity and sensitivity.

Scikit-learn package for python is equipped with tools to generate input for the ROC curve, which is TPR and FPR values for different thresholds. However, the final prediction is usually based on threshold equal to 0.5 if the decision function (a function assigning 0 or 1 based on the probability) has values in $[0, 1]$. Since the PIPPI study requires classifiers with sensitivity (TPR) and specificity (1-FPR) both above 0.7 [39], the decision function should be modified accordingly. Such models have a piece of the ROC curve in the area defined by $TPR \geq 0.7 \wedge FPR \leq 0.3$. If a model does not meet these conditions, a relaxed condition is applied, namely $TPR \geq 0.7 \wedge FPR \leq 0.4 \wedge \text{threshold} = \text{argmin}(FPR)$, or the default threshold is chosen. Otherwise, the best choice is the point in which ROC curve has derivative equal or less but closest to 1. This is because for such a point the exchange between TPR and FPR has no utility. If the derivative is bigger than 1, we can increase TPR more than FPR by moving the threshold down, otherwise there is an incentive to increase the threshold, reducing FPR more than TPR. If ROC function is differentiable, such point exists by the intermediate value theorem for the derivative or an adapted Rolle's theorem for ROC. Nonetheless it should be noted that there is no guarantee this point belongs to the area bounded by the conditions.

The relaxed condition allows for equal number of true positives and false positives if the proportion of target instances to non-target ones in the test set is 1:2 (which is almost the case here). This means that half of the positive assignments of the classifier cannot be trusted and for this reason a custom threshold was not defined for FPR lower than 0.4. An algorithm presenting this rationale can be found in section A.3 of the

appendix. The code works also with non-differentiable step functions. Since the models are to be compared in terms of AUC, it makes no change for this comparison to impose rules on the choice of the threshold. AUC takes all thresholds into account (as ROC is a collection of points computed from those thresholds), so it is indifferent of the final choice of one point on the curve.

Benchmark models are created to assess the quality of features. They are trained with the information about the patients' sex and age only. The models called *generic* apart from age and sex use the features created from EMR. Patterns are then added as an extension to those features. Because the number of patterns to mine is reduced to simplify computation, some information contained in rare features is lost, as they are not incorporated into any of the selected patterns. This is why these patterns should be used in addition to the non-temporal features and not instead of them. Moreover, for the same reason, some patients do not exhibit any of the frequent patterns. If the features were neglected, those patients would be represented by all-zero vectors and assigned the same prediction by the algorithms. An attempt was made to analyze models trained solely on patterns but they performed inconsiderably better than random. A more detailed description of how the patterns were chosen in this study can be found in section 4.1.

For further quantitative evaluation and comparison of the models, confusion matrices are exported together with ROC curves and calculated AUC and BIC. The best hyperparameters for each random-searched model are also saved for further tuning with GridSearch. For qualitative assessment feature importances (or coefficients for logit) are discussed with the medical experts.

4. Results

Presented below are the outcomes of preprocessing, mining patterns and learning from the dataset described in chapter 2. Section 4.1 contains aggregated information about the features and patterns, as well as parameters and settings that may serve to reproduce the output. The following sections analyze generic and enhanced models in both quantitative and qualitative way, and evaluate them according to section 3.4. The impact of patterns is discussed in section 4.4.

4.1. Experimental setup

After sanitization and feature engineering described in detail in 2.2 and 3.2, a matrix of 94093 rows and 3420 columns (see table 4.1) with 27.13% target response and 99.3% zeros was divided into age groups (see table 4.2). The intervals used for separating the instances were decided by the medical experts and reflect stages of children's life in the Dutch education system. A group of special interest are kids aged 4-11, as mental health problems tend to unveil when children attend elementary school [19, 60]. Indeed, the composition of groups 4-7 and 8-11 shows the highest percentage of target instances. Another reason for such segregation is lowering the variance of exposure for younger ages, i.e. a three-year-old patient can be registered only for 1, 2 or 3 years. For each of the groups training and testing sets were separated in proportion 7:3. The models were fitted with the training sets after dropping any columns that were generated by the test sets. It is worth noticing how the number of columns plummeted, which only confirms the sparsity of the matrix.

Table 4.1.: Number of features by source

source	features
ICPC	927
ATC	859
ATC3	87
tests (counts and times in between)	1783
referrals	100
free text	101
visits from *consultations*	1
personal data	2
deleted (appeared only in test sets)	- 440
total	**3420**

Source: own calculation

Table 4.2.: Composition by age

age	instances	features	% target in train	% target in test
0-3	15023	1217	21.07	20.68
4-7	23123	1861	30.32	31.99
8-11	18465	2112	36.63	37.62
12-15	13462	2180	30.05	30.75
16+	24010	2876	18.37	18.20

Source: own calculation

Events, which are also 1-patterns, were mined together for all patients allowing for computational efficiency but they were stored per patient and scored retaining age division. This is because frequency of occurrences vary with age – it is unlikely for a newborn to have an sexually transmitted disease [87] or for an adolescent to visit a GP to complain about subfebrile temperature or with no complaints at all [96]. Each pattern can therefore be unequally predictive for different groups. For the same reason there are different numbers of events and patterns mined for each age group. As expected, the older the patient, the more versatile symptoms can be reported (see table 4.4 for the number of distinct events). The number of 2-patterns per person also increases with age, which is presumably due to a longer observation time.

Table 4.3.: Number of distinct events by source

source	no. of events
ICPC	927
ATC3	87
tests	1247
referrals	100
free text	101
total	**2462**

Source: own calculation

Within each group, 300 best 1-patterns were chosen in total, 200 target-specific and 100 non-target-specific, which constituted to 15-33% of all events collected, depending on the group. These events became subpatterns for mining the 2-patterns using the algorithm described in section 3.3.5. Because of the aforementioned sparsity, combined supports of these patterns were not equal to the magnitudes of the samples, so some of the population was not covered by the 300 selected patterns and thus could not be covered by their superpatterns, as mentioned in section 3.5 Table 4.4 contains aggregated outcomes of the mining algorithms per age group: total number of distinct events, shares of the instances not covered by the 2-patterns and average number of 2-patterns for the remaining shares.

Table 4.4.: 2-patterns mined by age

age	instances	events	2-patterns	% without	2-patterns per inst.
0-3	15023	910	49127	36.71	48
4-7	23123	1308	93155	20.59	120
8-11	18465	1422	103781	17.64	163
12-15	13462	1532	114203	14.64	188
16+	24010	1999	162910	12.91	267

Source: own calculation

500 target-specific and 250 non-target specific 2-patterns per age group served as subpatterns for mining 3-patterns using algorithm A.2 and specificity of the 2-patterns again determined by the wilson score. The algorithm turned out to be fast – mining approximately 300.000 of 3-patterns on the basis of 750 2-patterns took less than 5 minutes for each age group. Table 4.5 presents the number of the 3-patterns per each group. Up to this point the selection of subpatterns was based on arbitrary numbers of specific patterns to control the scope of memory used. For choosing pattern extension to the matrix of non-temporal features it is more important to limit the quality of patterns rather than their number. Table 4.6 presents means and maximal values of the lower Wilson interval bounds by age group and pattern length. Wilson score is a modified probability (lowered for infrequent patterns) that a patient has a mental health problem, provided that the pattern is observed in their EMR. It comes as no surprise that 2-patterns experience a drop in average score, as they are less frequent than single events, and that for 3-patterns slightly higher scores are observed. This is due to the fact that they build up on the most frequent 2-patterns, pruning the support, which is the denominator of confidence in definition 3.14. Following this logic, it becomes clear that patterns of different lengths should not be applied the same rules for selection – otherwise the models could end up being trained only on target-specific 3-patterns and non-target-specific 2-patterns.

Table 4.5.: 3-patterns mined by age

age	3-patterns	% without	3-patterns per inst.
0-3	75866	71.07	66
4-7	112455	75.60	79
8-11	110719	77.01	61
12-15	125400	68.93	72
16+	142469	79.31	71

Source: own calculation

Table 4.6.: Average and maximum Wilson scores

age	Average scores			Maximum scores		
	1-patterns	2-patterns	3-patterns	1-patterns	2-patterns	3-patterns
0-3	0.1038	0.0784	0.1152	0.5325	0.6414	0.6778
4-7	0.1478	0.1061	0.1295	0.5748	0.6614	0.6778
8-11	0.1449	0.0970	0.1229	0.6110	0.6689	0.6778
12-15	0.0932	0.0616	0.0782	0.4620	0.5458	0.6230
16+	0.0712	0.0517	0.0643	0.3754	0.4741	0.5126

Source: own calculation

2-patterns achieving at least 0.50 and 3-patterns with scores equal to or above 0.55 were added as target-specific temporal features in age groups 0-3, 4-7, 8-11. For the two older groups thresholds of 0.40 and 0.45 were chosen for patterns of length 2 and 3 respectively, as their scores were significantly lower. The thresholds were arbitrary but not contingent. They were chosen by trial and error from the top of available Wilson score values to ensure qualitatively reasonable patterns. On the other hand, it is impossible to set such a threshold for non-target-specific patterns, as thousands of them have scores of 0. Their number was therefore set to be half of the number of target-specific patterns. Table 4.7 contains the results of the selection described above with thresholds in parentheses.

Table 4.7.: Number of features in models with patterns

age	target-specific		non-target-specific		total
	2-patterns	3-patterns	2-patterns	3-patterns	
0-3	29 (0.50)	48 (0.55)	14	24	1332
4-7	114 (0.50)	204 (0.55)	57	102	2338
8-11	100 (0.50)	119 (0.55)	50	59	2440
12-15	84 (0.40)	77 (0.45)	42	38	2421
16+	12 (0.40)	13 (0.45)	6	6	2913

Source: own calculation

The initial parameter grids did not vary between groups and were selected specifically for a binary classification task and considering the training set magnitude according to the knowledge of the packages' developers and previous experience. Depending on the attribute, the values sampled from an arithmetic or a logarithmic scale. The number of values tried for each hyperparameter was modest to enhance computational speed and the tuning process was repeated two to three times, each based on the output of the previous GridSearch. For SVM classifier linear and RBF kernels were tried with $C \in \{1, 2, 5\}$ and $\gamma \in \{0.01, 0.1, 1\}$. An automated (calculated by the package) value of γ was also used. The trees were allowed to grow until 4, 8 or 12 levels in depth, limiting the number of features tried for each split to 30%, 50% or square root of the total number of features available, or not limiting at all. Minimal samples required to generate a new split were set to 10, 100 and 500, and the splits were optimizing

either Gini information criterion or the entropy. The same parameters were imposed on trees in random forests, with additional assumption for the number of the trees in $\{10, 50, 100, 1000, 5000\}$, with or without bootstrap. Neural networks with 3, 4 or 5 hidden layers and 40, 50 or 100 neurons were run with learning rates 0.005 and 0.01. The algorithm was let to choose between ELU and $RELU$ as activation function and forced to stop after 20 or 30 checks without progress. Batch sizes of 64 or 128 were used with norm momentum 0.9 or no momentum imposed. Models with and without drop out rate (0.5) were tried. For XGBoost $\alpha \in \{0, 0.00001, 0.01, 0.1, 1, 10\}$ was tried in combinations with $\gamma \in \{0, 0.05, 0.1\}$ and learning rate 0.01, 0.03 and 0.05. Weight of 1 was fixed for positive instances as recommended for 2-class segregation and 50%, 80% or 90% of the sample was available for each tree. Other than that, the same tree characterizing parameters were searched from as in trees and random forests. The number of estimators for XGBoost was decided by running a cross-validated model with arbitrary parameters (learning rate 0.1, maximal depth of 5, $\gamma = 0$, 80% subsample of columns and instances by tree) and settling for the moment when the progress stopped. Logit, being the simplest of these methods, did not take any additional hyperparameters.

4.2. Generic models

To test the quality of the available data and effectiveness of feature engineering described in 3.2, models using only the non-temporal features were tried. This step is also fundamental to determine the impact of patterns. The results are presented in table 4.8 and figure 4.1 and a more detailed overview can be found in tables 4.10 and 4.11.

Most of the models managed to reach TPR$\geq$ 0.7 and FPR$\leq$ 0.4. The only models that failed to meet the relaxed condition in any of the groups were regression trees. They also achieved the worst AUC values or close to the lowest. This may be partially caused by the fact that ROC curve for trees is piecewise linear (it is hardly noticeable on the graph, as the trees were allowed to grow up to 12 nodes deep, resulting in $2^{12} = 4096$ leaves and therefore the same number of point on the curve), but most likely it is due to not having used all of the features. Logit and SVM performed moderately but firmly across all ages whereas deep neural networks (denoted by DNN) did rather badly on the youngest and the oldest group. The unquestionable winner in terms of AUC among the models trained on non-temporal data is XGBoost, achieving the desired specificity and sensitivity above 0.7 in age groups 4-7 and 8-11. It is slightly better than random forest, which is unsurprising, as both of them ensemble many regression trees to minimize the error.

Table 4.8.: AUC for generic models

age	logit	SVM	tree	RF	DNN	XGB
0-3	0.697287	0.714870*	0.685798	0.731793*	0.695566	**0.751310***
4-7	0.739850*	0.728655*	0.667533	0.738459*	0.756297*	**0.774178****
8-11	0.743810*	0.735697*	0.695151	0.756354*	0.759261*	**0.787335****
12-15	0.737930*	0.701949	0.668177	0.720493*	0.737212*	**0.761485***
16+	0.733091*	0.710279*	0.684641	0.744870*	0.639210	**0.761971***

** = met the conditions for specificity, sensitivity Source: own calculation

* = met the relaxed condition

bold = best AUC per age

Table 4.9.: AUC for benchmark models

age	logit	SVM	tree	RF	DNN	XGB
0-3	0.547847	0.564416	**0.567684***	**0.567684**	0.564416	0.560164
4-7	0.559554	0.515287	**0.585098***	**0.585098**	**0.585098**	**0.585098**
8-11	0.557195	0.545119	**0.561955***	0.560411	0.561542	0.561542
12-15	0.517042	0.512173	0.553687*	0.553406	**0.554031**	0.511683
16+	0.666845	0.551705	0.674179*	0.674179	**0.674762**	**0.674762**

* = assigned positives Source: own calculation

bold = best AUC per age

Benchmark models, except trees, failed to assign any of the test instances with target class, which means they identified all patients as healthy. As a result, the generic models performed better than the benchmark (table 4.9), on average making a difference of 0.18 in AUC between the best generic model and the best benchmark model for each group. However, for some classifiers, e.g. neural network in the group aged 16 and above, did worse than its equivalent trained only with age and sex. This can indicate that the model is overfitted in this particular case. Figure 4.1 consists of ROC curves for all five age groups. Plotted results are of the best models in each category – XGBoost for all generic models, random forests (0-3, 4-7), neural networks (12-15, 16+) and a tree (8-11) for benchmarks.

Table 4.10.: Output of the best generic models

age	sens.	spec.	confusion matrix[1]		% correct
0-3	0.7253	0.6526	2333 256	1242 676	66.76
4-7	0.7008	0.7022	3313 664	1405 1555	70.17
8-11	0.7054	0.7069	2443 614	1013 1470	70.63
12-15	0.7005	0.6764	1892 372	905 870	68.38
16+	0.7002	0.6877	4052 393	1840 918	69.00

Source: own calculation

Table 4.11.: Output of the best benchmark models

age	sens.	spec.	confusion matrix		% correct
0-3	0	1	3575 932	0 0	79.32
4-7	0	1	4718 2219	0 0	68.01
8-11	0.4745	0.4175	1816 870	1640 1214	54.69
12-15	0	1	2797 1242	0 0	69.25
16+	0	1	5892 1311	0 0	81.80

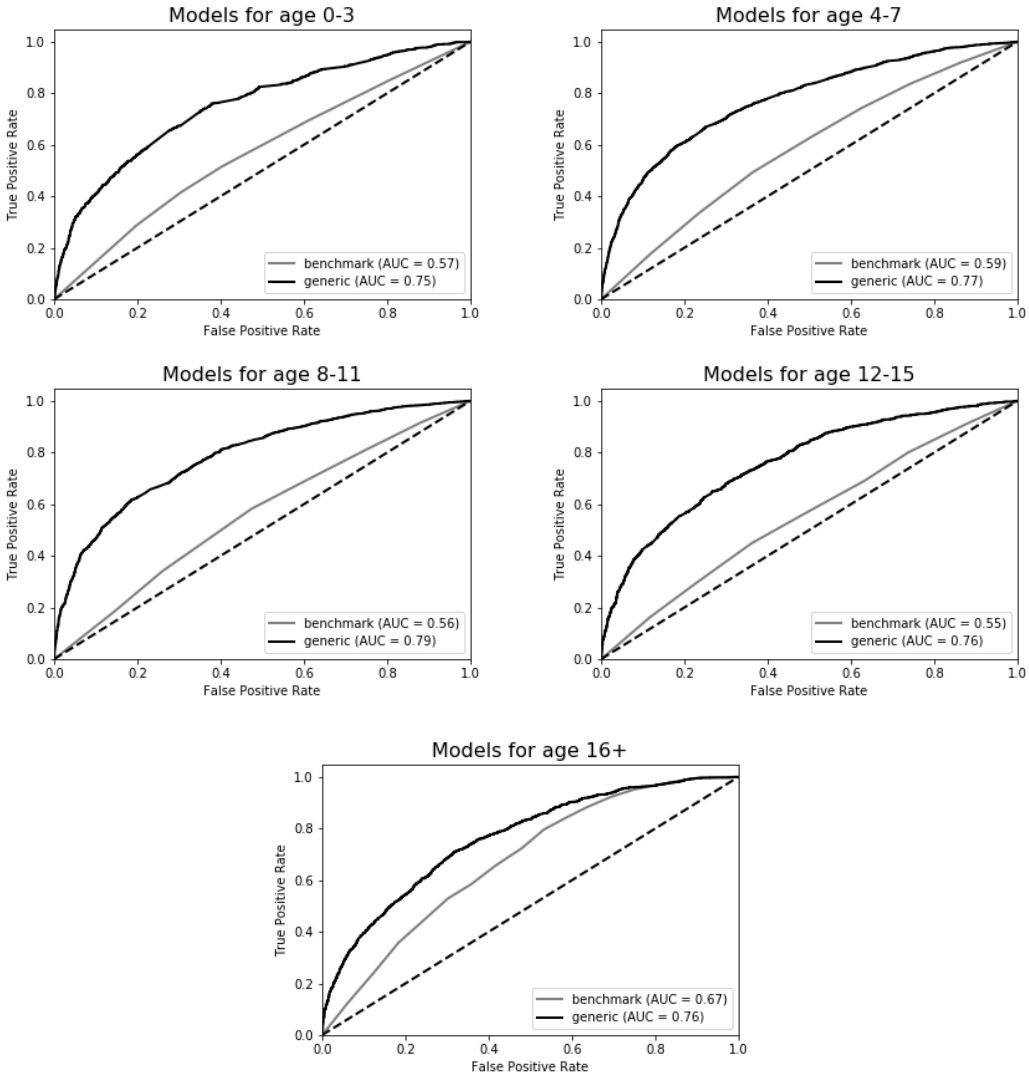
Source: own calculation

Qualitative evaluation of the features confirms that the number of visits is indeed of much impact when it comes to predicting mental issues, with the calculated importance at least twice as high as of the other features. In table 4.12 it is also evident that counts of all tests (`meetwaarden_count`) performed on patients are, as anticipated, of significant importance regardless of age group. Other aggregating features, namely the number of different tests (`meetwaarden_unique`), the frequency of having blood drawn (`blood_count`), as well as the minimal time between those events (`blood_min`) also appear as highly predictive. Other features that repeat across models include `ver_overig` (referral to *other*), `atc3_J01` (antibacterials [114]), especially `atc3_J01CA04`, which is

[1]How to read a confusion matrix:

		Prediction			
		Negative	Positive		
Reality	Negative	TN	FP	N = FP + TN	FPR = **FP** / N
	Positive	FN	TP	P = TP + FN	TPR = **TP** / P

Figure 4.1.: Receiver operating characteristic curves for generic and benchmark models



amoxicillin, a strong antibiotic [112]. ver_nan is difficult to decipher – it might stand for no referral or could be generated by an EMR system error at a general practice. If the latter is true, it might be beneficial for the models to remove this event from EMR. Another common feature, ver_huisartsgeneeskunde, denotes a referral to a GP, which could be advised during a consultation over a phone call, perhaps made by parents, as it only appears for ages 0-11. Other age-specific features selected by XGBoost models are: no illnes or upper respiratory tract illness for ages 0-3 (icpc_A97 and icpc_R74), ear inflammation for 4-7 (icpc_H71), medicine for respiratory system in the group of 8-11 year-olds (atc3_R06), referral to an X-ray in 12-15 (ver_r?ntgenologie) and contraceptives for young adults (atc_G03AA07). Medical experts agree that all of these feature importances make sense.

To examine whether a feature increases or decreases the probability of developing a mental illness, a tree graph could be investigated. However, in boosted models hundreds of trees are created, with up to 12 levels in depth, it would be therefore very difficult to infer the relation. Moreover, the relation might not be unambiguous – the same feature can be used a few times in the same tree, to split towards higher probability in one node and to lower it in another. A simpler idea to get the general direction of the effect a feature has is to export coefficients from logit models. The higher the positive value of a coefficient, the more target-specific a feature is. Analogically, negative coefficients indicate features correlated with healthy instances. Table 4.13 consists of signs of the coefficients for all features displayed in table 4.12. Even for features important for all models the signs vary across the groups, which implies that the decision to split the dataset by age is justified. An intriguing observation is that the relation of sex, denoted by M for *male*, with target variable is mostly positive (except ages 16+), which means that boys should be more prone to having mental diseases. This dependence was confirmed by other studies, where boys were more often identified with such problems in elementary school [29, 45, 65, 97, 110] but no association with gender was found in older children [12, 34, 73, 110] or younger [38, 71].

Table 4.12.: 10 most important features in XGBoost generic models by age group

0-3

feature	imp.[a]
visits_count	0.0952
ver_nan	0.0442
ver_overig	0.0367
age	0.0332
M	0.0250
atc3_J01	0.0248
icpc_R74	0.0241
ver_huisartsgeneeskunde	0.0233
atc_J01CA04	0.0203
icpc_A97	0.0185

4-7

feature	imp.
visits_count	0.0951
ver_overig	0.0353
age	0.0325
ver_nan	0.0266
atc3_J01	0.0230
meetwaarden_count	0.0224
ver_huisartsgeneeskunde	0.0179
atc_J01CA04	0.0177
icpc_R74	0.0155
icpc_H71	0.0140

8-11

feature	imp.
visits_count	0.0710
ver_overig	0.0318
age	0.0269
meetwaarden_count	0.0261
atc3_J01	0.0235
atc_J01CA04	0.0230
ver_nan	0.0230
atc3_D07	0.0131
ver_huisartsgeneeskunde	0.0124
atc3_R06	0.0122

12-15

feature	imp.
visits_count	0.0716
ver_overig	0.0277
meetwaarden_count	0.0253
atc3_J01	0.0228
age	0.0221
ver_nan	0.0201
atc_J01CA04	0.0162
ver_r?ntgenologie	0.0145
blood_min	0.0135
atc3_D07	0.0133

16+

feature	imp.
visits_count	0.0600
age	0.0321
meetwaarden_count	0.0304
ver_overig	0.0231
atc3_J01	0.0170
blood_min	0.0164
atc3_G03	0.0164
atc_G03AA07	0.0147
meetwaarden_unique	0.0147
blood_count	0.0126

Source: own calculation

[a] feature importance

Table 4.13.: Signs of selected coefficients in generic logit models

feature	0-3	4-7	8-11	12-15	16+
age	-	+	-	-	-
atc_G03AA07	n.a.[2]	-	-	+	-
atc_J01CA04	-	+	-	-	-
atc3_D07	-	-	-	-	-
atc3_G03	+	+	+	+	+
atc3_J01	+	-	+	-	-
atc3_R06	-	+	+	-	-
blood_count	+	+	+	+	-
blood_min	-	+	+	-	+
icpc_A97	-	-	-	-	-
icpc_H71	+	+	+	+	-
icpc_R74	+	-	-	-	-
M	+	+	+	+	-
meetwaarden_count	-	-	+	-	-
meetwaarden_unique	-	-	-	+	+
ver_huisartsgeneeskunde	-	-	-	-	-
ver_nan	-	+	-	-	+
ver_overig	-	-	-	-	-
ver_r?ntgenologie	+	-	+	+	+
visits_count	+	+	-	-	+

Source: own calculation

4.3. Models with patterns

Examples of events with the highest and lowest Wilson scores can be found in table 4.14. It can be seen that some events, e.g. icpc_D62 and icpc_H62, which both stand for administrative operations, repeat across age groups as the least target-specific. Although it is logical that they are unlikely to be associated with psychological issues, it is questionable whether such events should be allowed in the models. On the other hand, inc_GLUCB and dec_GLUCB, which mark increasing and decreasing glucose levels [113], may indicate diabetes, which some studies list as correlated with mental health, for instance Chatterton [15] recommends using it as a predictor for diabetes type 2 and Bădescu [3] claims depression occurrence is two to three times higher in people with diabetes. This shows that the scoring scheme is successful at finding predictors of proven quality.

[2]There was no record of atc_G03AA07 in this age group, hence it was not included in the model.

Table 4.14.: Most predictive events according to the Wilson score

age	best target-specific				best non-target-specific			
	event	train	target	score	event	train	target	score
0-3	atc3_A03	120	78	0.532	icpc_D62	10	0	0
	icpc_N88	10	9	0.492	icpc_H62	8	0	0
	met_CAMKDF	6	6	0.474	icpc_Y16	8	0	0
4-7	met_OD05BMD	9	9	0.575	icpc_D62	24	0	0
	inc_NEUTBMD	9	9	0.575	icpc_H62	13	0	0
	atc3_J05	37	29	0.574	icpc_A96	11	0	0
8-11	met_OW02DF	20	18	0.620	icpc_D62	13	0	0
	met_OW03DF	20	18	0.620	dec_QUETAO	12	0	0
	met_Q405B	46	36	0.596	met_YDB	11	0	0
12-15	inc_GLUCB	12	10	0.462	met_RDW	20	0	0
	atc3_H01	33	21	0.415	met_CAMPDF	18	0	0
	icpc_Z44	85	47	0.415	met_SHIPDF	18	0	0
16+	met_HOASRQ	4	4	0.375	inc_TRIGB	26	0	0
	inc_HBSIBMT	4	4	0.375	icpc_D62	18	0	0
	dec_GLUCB	60	26	0.283	icpc_R62	18	0	0

Source: own calculation

Table 4.15 presents frequent 2- and 3-patterns for each group. It can be observed that in some cases the most predictive events from table 4.14 form patterns that score even higher, e.g. `met_OW02DF-met_OW03DF` is built from target-specific events for ages 8-11. An interesting example of how longer patterns narrow down the variance of target predictiveness can be found in the group aged 0-3. 32 out of 38 patients who experienced `icpc_A03` (fever [114]) before being prescribed `atc3_A03` (drugs for gastrointestinal disorders, mostly antispasmodics) were identified with mental health problems, and all 14 of those who were also using `atc3_D02` (dermatological emolients) in between these events constituted to the target.

Table 4.15.: Most predictive patterns according to the Wilson score

age	best target-specific			
	pattern	train	target	score
0-3	icpc_A03-atc3_A03	38	32	0.641
	icpc_A03-atc3_D02-atc3_A03	14	14	0.678
4-7	atc3_J01-met_OW02DF	13	13	0.661
	met_OW02DF-met_O037DFMM-ver_overig	14	14	0.678
8-11	met_OW02DF-met_OW03DF	19	18	0.669
	atc3_R01-met_Q405B-met_Q502B	14	14	0.678
12-15	icpc_U35-icpc_A44	8	8	0.546
	icpc_H71-icpc_A44-icpc_L74	11	11	0.623
16+	icpc_H27-atc3_G03	6	6	0.474
	dec_GLUCB-dec_MCHB-inc_ERYBMT	7	7	0.513

age	best non-target-specific			
	pattern	train	target	score
0-3	icpc_A17-icpc_A97	20	0	0
	icpc_A62-angst-icpc_A80	14	0	0
4-7	pleeggezin-atc3_J01	21	0	0
	icpc_A15-icpc_R03-atc3_R03	16	0	0
8-11	icpc_S03-ver_neonatologie	30	0	0
	icpc_S03-ver_neonatologie-ver_overig	25	0	0
12-15	icpc_R78-met_ASHBRZ	28	0	0
	icpc_R78-met_ASHBRZ-atc3_R03	21	0	0
16+	ggz-atc3_R01	37	0	0
	ggz-atc3_R01-ver_overig	24	0	0

Source: own calculation

The same models were fitted with features enriched with patterns, retaining the training and testing sets of instances. Table 4.16 consists of respective AUC values calculated from performance on the test set. Unsurprisingly, XGBoost models have again achieved the highest scores within each age group. Table 4.17 illustrates the output of these models. It contains values for sensitivity (TPR) and specificity (1-FPR), together with full confusion matrices and calculated hit rate (the percentage of correct guesses on the test set). Taken everything into account, XGBoost in age group 8-11 seems to be the best model: it has the highest AUC, specificity and the biggest share of correct classifications. This might have to do with the fact that exposure to life events has a vast impact on school-aged kids [12, 110] but not toddlers [38, 71]. Even though not all the results met the expectations of the PIPPI project, the confusion matrices look satisfactory – in majority of models true positives outnumber any kind of mistakes (false positives or false negatives). The percentage of correct answers is similar across age groups and averages 69.4. Precise hyperparameters chosen in the tuning process can be found in table 4.18.

Table 4.16.: AUC for models with patterns

age	logit	SVM	tree	RF	DNN	XGB
0-3	0.698540	0.711909	0.688393	0.735659*	0.690844	**0.753800***
4-7	0.739943*	0.737361*	0.673611	0.740415*	0.761222*	**0.777084****
8-11	0.743869*	0.740431*	0.697411	0.762636*	0.760810*	**0.788060****
12-15	0.737957*	0.705690	0.668990	0.726506*	0.728695*	**0.758749***
16+	0.732934*	0.710509*	0.687669	0.747880*	0.640431	**0.763166***

** = met the conditions for specificity, sensitivity Source: own calculation

* = met the relaxed condition

bold = best AUC per age

Table 4.17.: Output of the best models with patterns

age	sens.	spec.	confusion matrix		% correct
0-3	0.7049	0.6831	2442 275	1133 657	68.76
4-7	0.7066	0.7020	3312 651	1406 1568	70.35
8-11	0.7006	0.7188	2484 624	972 1460	71.19
12-15	0.7013	0.6786	899 371	371 871	68.56
16+	0.7002	0.6775	3992 393	1900 918	68.17

Source: own calculation

Table 4.18.: Final hyperparameters in models with patterns

age	no. of trees	max. depth	learning rate	sample by tree	columns by tree	child weight	α	γ	λ
0-3	182	12	0.05	50%	60%	1	0.00001	0.04	1
4-7	281	12	0.04	70%	40%	1	0.05	0.05	1
8-11	223	12	0.05	70%	90%	1	0.00001	0.0	1
12-15	227	8	0.05	60%	90%	1	0.01	0.0	1
16+	265	9	0.04	80%	50%	1	0.0	0.075	1

Source: own calculation

Feature importance in the models with patterns (table 4.20) to a great extent repeated the outcome of generic models. The number of visits was again the main predictor and together with `age`, `ver_overig` and `atc3_J01` appeared in all age groups. Sex (denoted by `M`) was of significant importance in three models instead of one, diminishing the impact of `ver_huisartsgeneeskunde`. A few new events made it to the top 10, including medication from `R` (respiratory) and `D` (dermatological) categories [114]. Logit models, witnessing slightly lower AUC than XGBoost, rely on different features, as depicted in

table 4.19. It turns out that patterns played a more important role in linear models, it could therefore be a field worth exploring. Both tables confirm the aforementioned relation between sex and mental health in elementary school.

Table 4.19.: 5 most specific features in logit models with patterns

age	target-specific		non-target-specific	
	feature	coef.	feature	coef.
0-3	atc_D06BA01	1.0381	icpc_A62	-1.5608
	atc3_A06-atc3_D01-ver_nan	0.9483	atc3_J01-atc3_D07-icpc_T10	-1.1906
	icpc_R90	0.9370	icpc_S09	-0.9780
	icpc_Z25	0.8707	icpc_A69	-0.9143
	icpc_N07	0.8550	ver_med. microbiologie	-0.9100
4-7	icpc_H71-icpc_R78-atc3_D04	3.1738	ver_alg.maatschappelijk werk	-2.9076
	icpc_L81-icpc_R06	2.9257	icpc_D83-icpc_L14	-2.6825
	atc3_D04-ver_r?ntgenologie-atc3_R05	2.3893	icpc_A60	-2.3083
	icpc_B03-icpc_S18	2.2640	icpc_R62	-2.1601
	icpc_K80	2.2036	met_TEMPA000	-2.0995
8-11	icpc_R90	0.5136	icpc_A62	-0.8833
	icpc_A44	0.4775	ver_verloskunde (eerste lijn)	-0.6780
	M	0.4714	ver_gynaecologie	-0.6465
	icpc_H70	0.3922	met_RDWBMD	-0.5861
	icpc_N01	0.3725	ver_kindergeneeskunde	-0.4619
12-15	icpc_R75	0.4995	icpc_A62	-0.6766
	icpc_R90	0.4247	met_RDWBMD	-0.5176
	icpc_A44	0.4061	met_ASHBRZ	-0.3952
	icpc_R71	0.3658	ver_waarneming	-0.3516
	atc_R05DA04	0.3491	icpc_A99	-0.3291
16+	icpc_N02	0.4096	icpc_A62	-0.5102
	met_PABUB	0.4090	age	-0.3716
	ver_cesartherapie	0.3936	met_RDWBMD	-0.3530
	icpc_L97	0.3913	huisarts	-0.3340
	icpc_D93	0.3638	icpc_S06	-0.2880

Source: own calculation

Table 4.20.: 10 most important features in XGBoost models with patterns by age group

0-3

feature	imp.
visits_count	0.1232
ver_overig	0.0504
ver_nan	0.0439
age	0.0384
atc3_J01	0.0276
M	0.0251
icpc_R74	0.0244
ver_huisartsgeneeskunde	0.0240
atc_J01CA04	0.0182
icpc_H71	0.0180

4-7

feature	imp.
visits_count	0.0779
ver_overig	0.0384
ver_nan	0.0290
atc3_J01	0.0278
atc_J01CA04	0.0246
age	0.0241
meetwaarden_count	0.0195
atc3_R01	0.0169
atc3_D02	0.0164
atc3_D07	0.0162

8-11

feature	imp.
visits_count	0.0960
ver_overig	0.0356
age	0.0346
meetwaarden_count	0.0293
atc3_J01	0.0261
ver_nan	0.0228
atc_J01CA04	0.0225
icpc_R74	0.0142
atc3_R03	0.0139
M	0.0138

12-15

feature	imp.
visits_count	0.0868
meetwaarden_count	0.0349
ver_overig	0.0302
age	0.0285
atc3_J01	0.0250
ver_nan	0.0171
atc3_D07	0.0155
atc_J01CA04	0.0148
ver_r?ntgenologie	0.0146
M	0.0143

16+

feature	imp.
visits_count	0.0610
age	0.0327
meetwaarden_count	0.0253
ver_overig	0.0236
atc3_G03	0.0176
atc3_J01	0.0158
meetwaarden_unique	0.0155
blood_min	0.0151
blood_count	0.0142
atc_G03AA07	0.0132

Source: own calculation

4.4. Comparison

In juxtaposition with other mental health decision support systems, such as the aforementioned Multilayer Perceptron and Multiclass Classifier from 2017 [2], obtained AUC values are rather low. This is mainly due to the fact that the dataset is not directly related to psychosocial indicators and is very sparse.

With rare exceptions (logit for 16+, SVM for 0-3, DNN for 0-3, 12-15 and XGB for 12-15), AUC increased after adding patterns, by 0.003 on average, compared to generic models. Figure 4.3 presents ROC curves of the best models with patterns in comparison to the best generic models. The improvement generated by adding patterns is noticeable but low and the curves almost overlap.

Figure 4.3.: Receiver operating characteristic curves for models with patterns

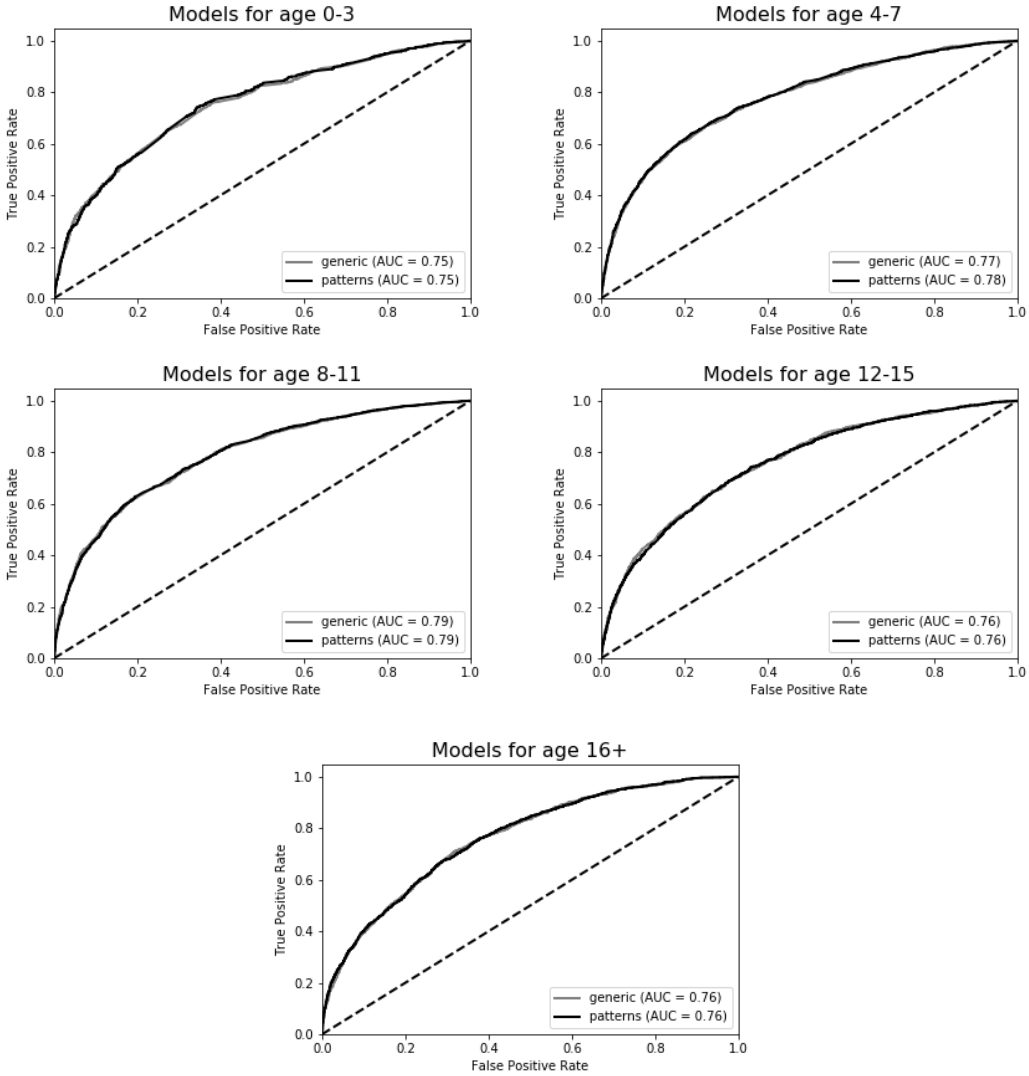


Table 4.21.: DeLong's statistical comparison of the ROC curves

age	negative instances	positive instances	AUC		z	p-value
			generic	patterns		
0-3	3575	932	0.751310	0.753800	-0.1796	0.8575
4-7	4718	2219	0.774178	0.777084	-0.3208	0.7484
8-11	3456	2084	0.787335	0.788060	-0.0774	0.9383
12-15	2797	1242	0.761485	0.758749	0.2219	0.8244
16+	5892	1311	0.761971	0.763166	-0.1043	0.9169

Source: own calculation

Two-tailed statistical tests [21] fail to confirm the difference between AUC as shown in table 4.21. There are a few possible explanations for the small change in AUC. First of all, the events alone seem already quite predictive and the generic models were extensively tuned, bringing out the best of the non-temporal data. Secondly, the patterns comprise a minority of the new matrices (1-20% of all features, 10.7% on average) and increase their sparsity, as they are far less common than single events. A rare target-specific pattern can make a change in a classifier's decision for the few patients who observed the pattern but it is less likely that it would influence the classification of the remaining patients. Another issue that might have occurred is that each patterns repeats information given in features and some other patterns, increasing correlation in the matrix, which is a rather unwanted phenomenon. It may cause the features to divest importance from the patterns. A possible solution to this would be to remove events that build a pattern when the pattern is added. A smart way to do it would be to compare the (conditional) score of a pattern given one or more of its components with scores of the components, then to make a decision of whether to include the pattern and drop features based on the calculated difference. There exists, however, the disadvantage of increasing the sparsity and hence decreasing the performance of the models. Finally, there is a chance that it is the events themselves, less their chronological order, that have impact on mental health. If so, no framework employing the timeline can add much predictive information to the non-temporal features.

Table 4.22.: BIC for benchmark models

age	logit	SVM	tree	RF	DNN	XGB
0-3	-4789.822	-4789.822	**-2254.742**	-4789.822	-4789.822	-4789.822
4-7	-3542.003	-3542.003	**-2371.208**	-3542.003	-3542.003	-3542.003
8-11	-1731.828	-1731.828	**-1089.034**	-1731.828	-1731.828	-1731.828
12-15	-2254.800	-2254.800	**-1380.100**	-2254.800	-2254.800	-2254.800
16+	-8837.166	-8837.166	**-3215.249**	-8837.166	-8837.166	-8837.166

Table 4.23.: BIC for generic models

age	logit	SVM	tree	RF	DNN	XGB
0-3	**4303.634**	6729.028	6843.237	6507.512	4510.950	6403.392
4-7	7003.484	7260.399	7725.859	6839.025	**6114.628**	6580.606
8-11	10974.385	11249.441	11674.877	10816.413	**10546.083**	10661.959
12-15	12728.190	**12034.502**	12848.162	12952.283	12601.595	12341.318
16+	14810.085	14972.944	16342.571	14759.922	15050.810	**14357.899**

Table 4.24.: BIC for models with patterns

age	logit	SVM	tree	RF	DNN	XGB
0-3	5199.615	5405.493*	7631.022	7419.344	5823.566	**4261.851***
4-7	11307.295	10959.649	11783.363	10977.044	10922.599	**10775.048**
8-11	13827.725	13991.797	14495.619	13721.591	13558.812	**13494.550**
12-15	14801.557	**13959.248**	14727.53	14867.858	14798.217	14555.800
16+	15305.462	15140.697	16555.488	**14742.876***	15459.706	14802.937

* = better than generic Source: own calculation

bold = best BIC per age

Models other than XGBoost emerged as the best in terms of Bayes Information Criterion. Predictably, the lowest BIC is observed for benchmark models. It has negative values, for the first term has a low multiplier (which is the number of features, i.e. 2) and the second term, $\frac{\mathcal{L}}{n}$, is close to or equals the ratio of target in test sample for each age group, which results in a negative value of logarithm. This is because all benchmark models but trees classified every instance as healthy. BIC values for models with patterns are higher than those of generic models by 1890.9 on average. Only three models managed to reach a lower value (marked with $*$ in table 4.24). Having failed to achieve a better score means that the additional information provided by the patterns was not enough to compensate for increased dimensionality. Improving the quality of patterns, as well as eliminating correlation could be good premises for lowering BIC.

5. Discussion

The study has developed two models that fulfill expectations of the PIPPI project and three propitious models that require a little improvement to meet them. Nevertheless, it is evident that mental health problems in young people can be successfully predicted with an AUC of over 0.75 using EMR from general practices. It has also been shown that temporal patterns created from this data can be used to enhance the performance of the models, both in terms of AUC and predictability.

The proposed methods of preprocessing EMR resulted in identifying the most important factors determining mental health, many of them featured in medical literature, which increases credibility of the models. The number of appointments with a GP has proved to be the main determinant, as mentioned in [73]. The hypothesis that the frequency of having laboratory tests performed would be a good predictor for mental health was based on this assumption and then confirmed by the feature importance. It has also been shown that gender influences the probability of having issues in the early ages, with boys being more vulnerable than girls. This phenomenon could be explained by the fact that they have higher problem rates in general and that mental health problems become more apparent when a child enters the school setting [19, 60].

The main contribution of this research is a framework for time and memory efficient pattern mining in a supervised setup, tailored to EMR intricacies and allowing for mental health prediction. A new method of scoring based on Wilson interval and assuming binomial distribution triumphed at identifying events and patterns that were both frequent and reasonable according to expert knowledge. The patterns occurred to be more useful for linear models than complex boosted estimators. The empirical part served to present the application of the newly developed framework and there are many opportunities to further test its efficiency in practice. From research perspective it would be worthy to compare the pattern scoring scheme with Batal's and other known methods on a different dataset – it is impossible using the EMR available to PIPPI, which was the reason for developing the modified approach.

Models were evaluated using AUC as required by the PIPPI project. The highest values per age group were achieved by XGBoost, which indicates that the relation of mental health problems and features acquired from EMR is non-linear. Models trained with temporal patterns accomplished better results than generic models in terms of AUC but not BIC. Percentage of correct classifications varied between 68.17% and 71.19%, reaching sensitivity of at least 70% and specificity above 67.75%.

This study has succeeded despite multiple hardships posed by the data. Overcoming these problems would create an opportunity to improve the result. The simplest idea would be to broaden the variety of Machine Learning techniques, adding more models or balancing the dataset, but the most promising action that could be taken to progress with

the research would be to implement more expert knowledge in data preprocessing. It is a well-documented fact that feature engineering can provide more gains than extensive altering of hyperparameters [5]. However, this cannot be easily done by machines and therefore is very time-consuming. A trade-off needs to be made between manual and Machine Learning ways of extracting information from the data.

The first step could be to qualitatively clean the EMR of all errors and typos. The data should be uniform, with one string of characters allowed to denote each level. Abbreviations, as observed in target-defining levels for referrals in table 3.1, widely increase disambiguation. Similarities between instances cannot be found if they are one-hot encoded as separate (and uncorrelated) features. As mentioned, it is also debatable whether events such as administrative operations should be included in the study, especially that they appear frequent. As they are all ICPC-coded, removing such records would require at least reading through (and, in this case, translating from Dutch) almost a thousand of code descriptions.

A few possible improvements have already been mentioned. A significant change could be observed if references for test results were available; it would allow for using numerical values and indicating whether they are too high or too low. Additional features indicating minimum, maximum and average levels could be incorporated. Exploring child health literature, e.g. [98], and creating a rule-based reference system for the test results is a potential direction in further development of a risk-identifying tool.

Another missing reference is the link between tables containing *consultations* and *episodes*. Being able to connect visits to the episodes of a disease would enable counting those visits per episode or per disease. Since the number of visits has proven to be a good predictor for mental health issues, the number of visits where a predictive symptom was raised has a high chance of being a good proxy as well. Moreover, the models used in this study operate on numerical variables, so using a variety of values instead of binary levels could possibly enhance their performance, as they would be able to measure distance between instances with more accuracy [31].

Medical knowledge could also be applied to enrich the data. Adding the most common side effects of medication or symptoms accompanying diagnosed diseases (e.g. tuberculosis and cough, broken leg and leg pain) would help the sparsity of matrices. Furthermore, these events could potentially be added with weights reflecting the probability of their appearance, adding desired variation to the otherwise binary features.

A less arbitrary approach to grouping medication could enhance the predictability as well. It might not only capture the similarities in composition or the area of impact, but also reflect the occasion of having it prescribed or the side effects. This might especially impact the studies of mental health, as taking some kinds of medication often results in malaise and may affect psychological condition in the long run. Prospective unsupervised methods for different clustering of medication include Principal Component Analysis and K-means.

The assumptions for mining patterns, i.e. the number of target-specific or non-target-specific subpatterns and the thresholds for deciding which patterns should be added as features, were made arbitrarily by trial and error. It remains to be explored if allowing for more various subpatterns would enhance the performance of the models, although it

is likely that events which scored lower on the Wilson scale would produce lower quality patterns. A different mixture of pattern lengths could be used, employing the most predictive of 4-patterns and 5-patterns mined. Another aspect worth considering is that patterns witness a nested structure: if $\mathcal{P}$ is frequent, all instances covered by $\mathcal{P}$ are covered by all of its subpatterns, which are also a result of the mining method. This causes a problem which Batal calls *spurious patterns*: patterns that are evaluated as predictive, yet redundant if given one of their subpatterns [4]. Such patterns increase correlation between features but do not enhance performance of models. This issue has not been resolved in this study and could become another area of improvement in the future.

Given the criticism of AUC as a method of scoring in section 3.4.1, a more objective-oriented measure could be proposed. An ideal solution to defining the evaluation metric would be to precisely quantify the economical and social costs of making each type of mistakes by the risk identification tool and include them in a custom-made loss function. This function would then serve both to compare between best-performing classifiers and as a scoring function to tune hyperparameters of the models. This would, however, require a separate and thorough study, and would depend on the country and its health-care system, making the tool only locally applicable. A simple alternative would be to set a fixed arbitrary weight between false negatives and false positives. Currently it is dependent on the proportion of target instances in the sample to non-target ones and equals approximately 1:2, meaning that decreasing the number of false negatives by 1 and adding 2 false positives instead is as good as no change to AUC. Such ratio could be decided by medical experts, reflecting a desired trade-off. Similarly BIC could be altered, as the sum of squares treats different types of mistakes equally.

Another idea would be to neglect possible thresholds for decision functions of the algorithms completely, together with their respective values of TPR and FPR, concentrating only on the chosen threshold and the outcome of its application. Ultimately, it is not the threshold that could have been chosen, but the one settled for, that influences the performance of a tool developed for the PIPPI project. Hence, the focus could alternatively be shifted to the confusion matrix and a loss function could be built based on its four components.

Perhaps it would also be worthwhile to consider differentiating penalties across ages. It could be a sensible strategy to have a classifier that is conservative for early ages and alerts only when it is very likely that the patient has mental issues, but sensitive for any signals in adolescents. Firstly, a child who passes as healthy in a restrictive algorithm at the age of 3 would likely come back to see their GP in the following years, possibly adding symptoms to the medical history that would help the algorithm recognize them as the target with more certainty, or on the contrary: more non-target-specific EMR would be added, advocating for the child's mental stability. Secondly, such approach would avoid putting a child and their family under unnecessary pressure in case the algorithm made a wrong decision, as it could be stressful for the child to receive mental treatment at young age [54]. On the other hand, young adults could conceivably cope with such diagnosis better, as their cognitive abilities would allow for an improved understanding. They can also decide for themselves whether to undergo such treatment, as opposed to

the minors, for whom the call is made by their guardians.

Of course, since the aim of the PIPPI project is to develop a decision supporting device, it would still be in line with this strategy to make the tool equally sensitive for all ages and let the GP conclude if any action should be taken. The action can be anything between an extended interview and therapy referral. Screening tools can increase awareness in professionals and decrease variation in their diagnoses but might also cause an increase in the number of false positives [84].

Bibliography

[1] Allen F. *Towards a general theory of action and time.* Artificial Intelligence, 23, 1984. p. 123-154.

[2] Anujume et al. *Performance Analysis of Machine Learning Techniques to Predict Mental Health Disorders in Children.* International Journal of Innovative Research in Computer and Communication Engineering, 5(5), 2017.

[3] Bădescu at al. *The association between Diabetes mellitus and Depression.* Journal of Medicine and Life, 2016.

[4] Batal et al. *A Temporal Pattern Mining Approach for Classifying Electronic Health Record Data.* ACM Trans Intell Syst Technol, 2013.

[5] Bergstra et al. *Algorithms for Hyper-Parameter Optimization.* Advances in Neural Information Processing Systems, 24, 2011.

[6] Berndt et al. *Health Care Use And At-Work Productivity Among Employees With Mental Disorders.* Health Affairs, 19(4), 2000. p. 244-256.

[7] Bezem et al. *A novel triage approach of child preventive health assessment: an observational study of routine registry-data.* BMC health services research, 14(1), 2014. p. 498.

[8] Bijl et al. *The prevalence of treated and untreated mental disorders in five countries.* Health Aff, 22(3), 2003. p. 122-133.

[9] Brown J.D., Riley A.W., Wissow L.S. *Identification of youth psychosocial problems during pediatric primary care visits.* Administration and Policy in Mental Health and Mental Health Services Research, 34(3), 2007. p. 269-281.

[10] Bruce B.G., Edward S.H. *Rule-based Expert Systems: The MYCIN Experiments of the Stanford Heuristic Proramming Project.* Addison-Wesley Reading, 1984.

[11] Bruce et al. *Coping with Chronic Illness in Childhood and Adolescence.* Annual Review of Clinical Psychology, 8, 2012. p. 455-480.

[12] Brugman et al. *Identification and management of psychosocial problems by preventive child health care.* Archives of Pediatrics & Adolescent Medicine, 2001.

[13] Cannon D.S., Allen S.N. *A comparison of the effects of computer and manual reminders on compliance with a mental health clinical practice guideline*. Journal of the American Medical Informatics Association, 7, 2000. p. 196-203.

[14] Cestnik B., Kononenko I., Bratko I. *ASSISTANT 86 : A knowledge elicitation tool for sophisticated users*. Sigma Press, 1987.

[15] Chatterton et al. *Risk identification and interventions to prevent type 2 diabetes in adults at high risk: summary of NICE guidance*. BMJ, 2012.

[16] Chen T., Guestrin P. *XGBoost: A Scalable Tree Boosting System*. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016. p. 785-794.

[17] Clancey W.J., Shortliffe E.H., Buchanan B.G. *Intelligent computer-aided instruction for medical diagnosis*. Proceedings of the Annual Symposium on Computer Application in Medical Care, American Medical Informatics Association, 1979.

[18] Collishaw et al. *Time trends in adolescent mental health*. The Journal of Child Psychology and Psychiatry, 2008.

[19] Crone M.R., Zeijl E., Reijneveld S.A. *When do parents and child health professionals agree on childs psychosocial problems? Cross-sectional study on parentchild health professional dyads*. BMC psychiatry, 16(1), 2016. p. 151.

[20] Cunningham et al. *Williams obstetrics, 25th edition*. McGraw-Hill, 2018.

[21] DeLong E., DeLong D., Clarke-Pearson D. *Comparing the areas under two or more correlated receiver operatingcharacteristic curves: a nonparametric approach*. Biometrics, 1988. p. 837-845.

[22] Garralda E. *Child and adolescent psychiatry in general practice*. Australian and New Zealand Journal of Psychiatry, 35(3), 2001. p. 308-314.

[23] Gerstman, B. *Basic Biostatistics: Statistics for Public Health Practice*. Jones & Bartlett Publishers, 2008.

[24] Goodman A., Joyce R., Smith J.P. *The long shadow cast by childhood physical and mental problems on adult life*. Proceedings of the National Academy of Sciences of the United States of America, 2011.

[25] Harris D., Harris S. *Digital design and computer architecture* Morgan Kaufmann, 2012. p. 129.

[26] Hofstra M.B., van der Ende J., Verhulst F.C. *Child and adolescent problems predict DSM-IV disorders in adulthood: a 14-year follow-up of a Dutch epidemiological sample*. Journal of the American Academy of Child and Adolescent Psychiatry, 41(2), 2002. p. 182-189.

[27] Hoogendoorn M., Burkhardt F. *Machine Learning for the Quantified Self: On the Art of Learning from Sensory Data*. Springer, 2018. p. 54-56.

[28] Hoppner F. *Knowledge discovery from sequential data*. Technical University Braunschweig, 2003.

[29] Horwitz S.M., Leaf P.J., Leventhal J.M. *Identification of psychosocial problems in pediatric primary care - Do family attitudes make a difference?* Archives of Pediatrics & Adolescent Medicine, 152(4), 1998. p. 367-371.

[30] Horwitz et al. *Identification and management of psychosocial and developmental problems in community-based, primary care pediatric practices* Pediatric, 89(3), 1992. p. 480-485.

[31] Janecek et al. *On the Relationship Between Feature Selection and Classification Accuracy*. Journal of Machine Learning Research, 4, 2008. p. 90-105.

[32] shan Kam P., chee Fu A.W. *Discovering temporal patterns for interval-based events*.Proceedings of the International Conference on Data Warehousing and Knowledge Discovery (DaWaK), 2000.

[33] Kass R.E., Raftery A.E. *Bayes Factors*. Journal of the American Statistical Association, 90(430), 1995. p. 773795.

[34] Kelleher et al. *Patient race and ethnicity in primary care management of child behavior problems: A report from PROS and ASPN*. Medical Care, 37(11), 1999. p. 1092-1104.

[35] Keyes et al. *The burden of loss: unexpected death of a loved one and psychiatric disorders across the life course in a national study*. The American journal of psychiatry, 171(8), 2014. p. 864871.

[36] Kieling et al. *Child and adolescent mental health worldwide: evidence for action*. The Lancet, 278(9801), 2011. p. 1515-1525.

[37] Kim-Cohen et al. *Prior juvenile diagnoses in adults with mental disorder: developmental follow-back of a prospective-longitudinal cohort*. Archives of general psychiatry, 60(7), 2003. p. 709-717.

[38] Klein et al. *Identification and management of psychosocial problems among toddlers by preventive child health care professionals*. European journal of public health, 20(3), 2010. p. 332-338.

[39] Koning N.R., Buchner F.L., Crone M.R. *Primary care integrated for identification of psychosocial problems in children (PIPPI)*. Research Protocol at Leiden University Medical Center, Afdeling Public Health en Eerstelijnsgeneeskunde, 20979, 2016.

[40] Kononenko I. *Semi-naive Bayesian classifier*, Proceedings of European Working Session on Learning, Springer Verlag, 1991. p. 206-219.

[41] Kononenko I. *Estimating attributes: Analysis and extensions of RELIEF*. Proceedings of European Conference on Machine Learning, Springer Verlag, 1994. p. 171-182.

[42] Kononenko I. *Machine Learning for Medical Diagnosis: History, State of the Art and Perspective*. Artificial Intelligence in Medicine, 23(1), 2001. p. 89-109.

[43] Kuryati et al. *Investigating machine learning techniques for detection of depression using strctural MRI volumetric features*. International journal of bioscience, biochemistry and bioinformatics, 3, 2013. p. 444-448.

[44] Kramer T., Garralda M.E. *Child and adolescent mental health problems in primary care*. Advances in Psychiatric Treatment, 6(4), 2000. p. 287-294.

[45] Leaf et al. *Pediatricians' training and identification and management of psychosocial problems*. Clinical Pediatrics, 43(4), 2004. p. 355-365.

[46] Levenson et al. *Social Media Use Before Bed and Sleep Disturbance Among Young Adults in the United States: A Nationally Representative Study*. Sleep, 40(9), 2017.

[47] Lewis et al. *Computerized assessment of common mental disorders in primary care: effect on clinical outcome*. Family Practice, 13, 1996. p. 120-126.

[48] van Lier et al. *Which better predicts conduct problems? The relationship of trajectories of conduct problems with ODD and ADHD symptoms from childhood into adolescence*. Journal of Child Psychology and Psychiatry, 48(6), 2007. p. 601-608.

[49] Loveday et al. *Pattern Recognition as an Indicator of Diagnostic Expertise* [in] Advances in Intelligent Systems and Computing. Springer, 2013.

[50] Maimon O., Rokach L. *Data Mining and Knowledge Discovery Handbook*. Springer, 2009. p. 875-877.

[51] Martinez R., Reynolds S., Howe A. *Factors that influence the detection of psychological problems in adolescents attending general practices*. British Journal of General Practice, 56(529), 2006. p. 594-599.

[52] Mehdibet al. *Data mining approaches for genome-wide association of mood disorders*. Psychiatric genetics, 22, 2012. p. 55-61.

[53] Michie D., Spiegelhalter D.J., Taylor C.C. *Machine learning, neural and statistical classification*, Ellis Horwood, 1994.

[54] Moses T. *Stigma and Self-Concept Among Adolescents Receiving Mental Health Treatment*. American Journal of Orthopsychiatry, 2010.

[55] Moskovitch R., Shahar Y. *Medical temporal-knowledge discovery via temporal abstraction* Proceedings of the American Medical Informatics Association (AMIA), 2009.

[56] Murray et al. *Disability-adjusted life years (DALYs) for 291 diseases and injuries in 21 regions, 1990-2010: a systematic analysis for the Global Burden of Disease Study 2010.* The Lancet, 380(9859), 2012. p. 2197-2223.

[57] Nielen et al. *Berekening morbiditeitscijfers op basis van NIVEL Zorgregistraties.* NIVEL, 2016. p. 12.

[58] Nugent et al. *Probability theory in the diagnosis of Cushing's syndrome.*The Journal of Clinical Endocrinology and Metabolism 24(7), 1964. p. 621-627.

[59] Olders H. *Average sunrise time predicts depression prevalence.* Journal of Psychosomatic Research, 55, 2003. p. 99105.

[60] Olfson M., Druss B.G., Marcus S.C. *Trends in mental health care among children and adolescents.* New England Journal of Medicine, 372(21), 2015. p. 2029-2038.

[61] Ormel et al. *Mental health in Dutch adolescents: a TRAILS report on prevalence, severity, age of onset, continuity and co-morbidity of DSM disorders.* Psychological Medicine, 2014. p. 1-16.

[62] Pantic I. *Online Social Networking and Mental Health.* Cyberpsychology, behavior, and social networking, 17(10), 2014.

[63] Pantic et al. *Association between online social networking and depression in high school students: behavioral physiology viewpoint.* Psychiatria Danubina, 24, 2012. p. 9093.

[64] Papapetrou et al. *Discovering frequent arrangements of temporal intervals.* Proceedings of the International Conference on Data Mining (ICDM), 2005.

[65] Park J.H., Bang Y.R., Kim C.K. *Sex and Age Differences in Psychiatric Disorders among Children and Adolescents: High-Risk Students Study.* Psychiatry Inverstigation, 2014.

[66] Penckofer et al. *Vitamin D and Depression: Where is all the Sunshine?* Issues in mental health nursing., 31(6), 2010. p. 385-393.

[67] Qiu et al. *Electronic Health Record Driven Prediction for Gestational Diabetes Mellitus in Early Pregnancy.* Nature research, 2017.

[68] Ragavan H., Rendell L. *Lookahead feature construction for learning hard concepts.* Proceedings of the 10th International Conference on Machine Learning, Morgan Kaufmann, 1993. p.252-259.

[69] Reddy C., Aggrawal C. *Healthcare Data Analytics*. CRC Press, 2015. p. 379-402.

[70] Reijneveld et al. *Area deprivation and child psychosocial problems - A national cross-sectional study among school-aged children*. Social Psychiatry and Psychiatric Epidemiology, 40(1), 2005. p. 18-23.

[71] Reijneveld et al. *Identification and management of psychosocial problems among toddlers in Dutch preventive child health care*. Archives of Pediatrics & Adolescent Medicine, 158(8), 2004. p. 811-817.

[72] Reijneveld et al. *Psychosocial problems among immigrant and nonimmigrant children - Ethnicity plays a role in their occurrence and identification*. European Child & Adolescent Psychiatry, 14(3), 2005. p. 145-152.

[73] Richardson et al. *Factors Associated with Detection and Receipt of Treatment for Youth with Depression and Anxiety Disorders*. Academic Pediatrics, 10(1), 2010. p. 36-40.

[74] Rink B., Harabagiu S., *Determining Relational Similarity Using Lexical Patterns*. Proceedings of the First Joint Conference on Lexical and Computational Semantics (*SEM), Association for Computational Linguistics, 2012. p. 413418.

[75] Rollman et al. *A randomized trial using computerized decision support to improve treatment of major depression in primary care*. Journal of General Internal Medicine, 17, 2002. p. 493-503.

[76] Rozita et al. *Employing artificial intelligence techniques in Mental Health Diagnostic Expert System*. Proceedings of In Computer & Information Science International Conference, 2012. p. 495-449.

[77] Rumelhart D.E., Hinton G.E. *Learning internal representations by error propagation*. MIT Press, 1986.

[78] Safer D.J. *Is ADHD Really Increasing in Youth?*. Journal of Attention Disorders, 22(2) 2015. p. 107-115.

[79] Sayal K. *Annotation: Pathways to care for children with mental health problems*. Journal of Child Psychology and Psychiatry, 47(7), 2006. p. 649-659.

[80] Sayal K., Taylor E. *Detection of child rental health disorders by general practitioners* British Journal of General Practice, 54(502), 2004. p. 348-352.

[81] Scholle et al. *Physician gender and psychosocial care for children: Attitudes, practice characteristics, identification and treatment*. Medical Care, 39(1), 2001. p. 26-38.

[82] Schriger et al. *Enabling the diagnosis of occult psychiatric illness in the emergency department: a randomized, controlled trial of the computerized, selfadministered PRIME-MD diagnostic system.* Annals of Emergency Medicine, 37, 2001. p. 132-140.

[83] Shahar Y. *A Framework for Knowledge-Based Temporal Abstraction.* Artificial Intelligence, 90, 1997. p. 79133.

[84] Sheldrick R.C., Merchant S., Perrin E.C. *Identification of Developmental-Behavioral Problems in Primary Care: A Systematic Review.* Pediatrics, 128(2), 2011. p. 356-363.

[85] Sherman et al. *The Power of the Like in Adolescence: Effects of Peer Influence on Neural and Behavioral Responses to Social Media.* Psychological Science, 27(7), 2016.

[86] Sidani et al. *The Association between Social Media Use and Eating Concerns among US Young Adults.* Journal of the Academy of Nutrition and Dietetics, 116(9), 2016. p. 14651472.

[87] Stratton K.R., Durch J.S., Lawrence R.S. *Vaccines for the 21st Century: A Tool for Decisionmaking.* The National Academies Press, 2000.

[88] Thorley C. *Not by degrees: Improving student mental health in the UK's universities.* Institute for Public Policy Research, 2017.

[89] Tick N.T., van der Ende J., Verhulst F.C. *Ten-year increase in service use in the Dutch population.* European child & adolescent psychiatry, 17(6), 2008. p. 373-380.

[90] Trivedi s., Pardos Z. Heffernan N.T. *Clustering Students to Generate an Ensemble to Improve Standard Test Score Predictions.* Springer, 2011.

[91] Twenge et al. *Increases in Depressive Symptoms, Suicide-Related Outcomes, and Suicide Rates Among U.S. Adolescents After 2010 and Links to Increased New Media Screen Time.* Clinical Psychological Science, 6(1), 2017.

[92] Tylee et al. *Youth-friendly primary-care services: how are we doing and what more needs to be done?* The Lancet, 369(9572), 2007. p. 1565-1573.

[93] Ûstùn T.B., Sartorius N. *Mental Illness in General Health Care: An International Study.* John Wiley & sons, 1995.

[94] Valko M., Hauskrecht M. *Feature importance analysis for patient management decisions.* Proceedings of medical informatics (MedInfo), 2010.

[95] Vallance et al. *Managing child and adolescent mental health problems in primary care: taking the leap from knowledge to practice.* Primary health care research & development, 12(4), 2011. p. 301-309.

[96] Viner R. *ABC of Adolescence*. Blackwell Publishing Ltd, 2005. p. 13-16.

[97] Vogels et al. *Identification of children with psychosocial problems differed between preventive child health care professionals*. Journal of Clinical Epidemiology, 61(11), 2008.

[98] de Vries et al. *Laboratoriumdiagnostiek bij kinderen: een praktische handleiding*. Prelum Uitgevers, 2015.

[99] Warner H.R. *Computer-Assisted Medical Decision-Making*. Academic Press Inc, 1979.

[100] Wiefferink et al. *Screening for psychosocial problems in 56-year olds: a randomised controlled trial of routine health assessments*. Patient education and counseling, 60(1), 2006. p. 57-65.

[101] Wieske et al. *Preventive youth health care in 11 European countries: an exploratory analysis*. International journal of public health, 57(3), 2012. p. 637-641.

[102] Wildman B.G., Kizilbash A.H., Smucker W.D. *Physicians' attention to parents' concerns about the psychosocial functioning of their children*. Archives of Family Medicin,8(5), 1999. p. 440-444.

[103] Wilson E.B. *Probable inference, the law of succession, and statistical inference*. Journal of the American Statistical Association, 22, 1927. p. 209212.

[104] Wittchen H.U., Mhlig S., Beesdo K. *Mental disorders in primary care*. Dialogues in Clinical Neuroscience, 5(2), 2003. p. 115128.

[105] Wittchen H.U., Nelson C.B., Lachner G. *Prevalence of mental disorders and psychosocial impairments in adolescents and young adults*. Psychol Med, 28(1), 1998. p. 109-126.

[106] Winarko E., Roddick JF. *Armada an algorithm for discovering richer relative temporal association rules from interval-based data*. Data and Knowledge Engineering, 63, 2007. p. 7690.

[107] Yap R.H., Clarke D.M. *An expert system for psychiatric diagnosis using the DSM-III-R, DSM-IV and ICD-10 classifications*. Proceedings of the AMIA Annual Fall Symposium, 1996. p. 229233.

[108] Zimmermann A., Nijssen S. *Supervised Pattern Mining and Applications to Prediction* [in] *Frequent pattern mining*. Springer, 2014. p. 425-442.

[109] Zwaanswijk M. *Pathways to Care: Help-seeking for child and adolescent mental health problems*, Utrecht University, 2005.

[110] Zwaanswijk et al. *Consultation for and identification of child and adolescent psychological problems in Dutch general practice*. Family Practice, 2005.

[111] Zwaanswijk et al. Help seeking for emotional and behavioural problems in children and adolescents: a review of recent literature. European Child & Adolescent Psychiatry, 12(4), 2003. p. 153-161.

[112] *Amoxicillin*. The American Society of Health-System Pharmacists Monograph. https://www.drugs.com/monograph/amoxicillin.html

[113] *Bepalingenclusters by Nederlands Huisartsen Genootschap*, 2015. https://referentiemodel.nhg.org/tabellen/inkijkexemplaren

[114] *The Anatomical Therapeutic Chemical Classification System with Defined Daily Doses (ATC/DDD)*. http://www.who.int/classifications/atcddd/en/

[115] *Waardelijst ICPC-1-2000NL 20111012*. https://decor.nictiz.nl/ketenzorg/kz-html-20141013T173536/voc-2.16.840.1.113883.2.4.3.11.60.103.11.12-2011-10-12T000000.html

[116] *WHO Collaborating Centre for Drug Statistics Methodology*https://www.whocc.no/atc/structure_and_principles

A. Code

All of the following snippets are coded in python 3.5.0.

A.1. Algorithm mining 2-patterns

NB. After mining the patterns, it is computationally important to keep only the earliest records per patient and pattern (e.g. by taking the mininum of the date).

```
def create_2_pat(pat,events,good,path,filename):
    ev_cols = ['event','date']
    events_pat = events[ev_cols].loc[events['pseudopatnummer'] \
                == pat]
    for ev in events_pat.iterrows():
        if ev[0] == 0:
            date = ev[1]['date']
            prev_all = [ev[1]['event']]
            prev_date_all = [ev[1]['event']]
            if ev[1]['event'] in good:
                prev_good = [ev[1]['event']]
                prev_date_good = [ev[1]['event']]
            else:
                prev_good = []
                prev_date_good = []
        elif ev[1]['date'] == date:
            temp = pd.DataFrame(columns = \
                        ['pseudopatnummer','ev_1','ev_2','end'])
            if ev[1]['event'] in good:
                temp['ev_1'] = prev_all + \
                                [ev[1]['event']] * len(prev_date_all)
                temp['ev_2'] = [ev[1]['event']] * len(prev_all) \
                                + prev_date_all
                temp['end'] = ev[1]['date']
                temp['pseudopatnummer'] = pat
                temp = temp.loc[temp['ev_1'] != temp['ev_2']]
                temp.to_csv(path+filename,mode='a',header=False, \
                prev_good = list(set(prev_good + [ev[1]['event']]))
                prev_date_good = list(set(prev_date_good \
                                + [ev[1]['event']]))
```

```
            else:
                if len(prev_good) != 0:
                    if len(prev_date_good) != 0:
                        temp['ev_1'] = prev_good \
                                       + [ev[1]['event']] \
                                * len(prev_date_good)
                        temp['ev_2'] = [ev[1]['event']] \
                                       * len(prev_good) \
                                       + prev_date_good
                    else:
                        temp['ev_1'] = prev_good
                        temp['ev_2'] = ev[1]['event']
                    temp['end'] = ev[1]['date']
                    temp['pseudopatnummer'] = pat
                prev_all = list(set(prev_all + [ev[1]['event']]))
                prev_date_all = list(set(prev_date_all \
                            + [ev[1]['event']]))
        else:
            temp = pd.DataFrame(columns = \
                        ['pseudopatnummer','ev_1','ev_2','end'])
            if ev[1]['event'] in good:
                temp['ev_1'] = prev_all
                temp['ev_2'] = ev[1]['event']
                temp['end'] = ev[1]['date']
                temp['pseudopatnummer'] = pat
                temp = temp.loc[temp['ev_1'] != temp['ev_2']]
                temp.to_csv(path+filename,mode='a',header=False, \
                                index=False)
                prev_good = list(set(prev_good + [ev[1]['event']]))
                prev_date_good = [ev[1]['event']]
            else:
                if len(prev_good) != 0:
                    temp['ev_1'] = prev_good
                    temp['ev_2'] = ev[1]['event']
                    temp['end'] = ev[1]['date']
                    temp['pseudopatnummer'] = pat
                    temp.to_csv(path+filename,mode='a', \
                                header=False,index=False)
                prev_date_good = []
            prev_all = list(set(prev_all + [ev[1]['event']]))
            prev_date_all = [ev[1]['event']]
            date = ev[1]['date']
    return None
```

A.2. Algorithm mining patterns of length 3 and more

The length of the patterns depends on the length of the subpatterns passed to the function as `patterns`. The following code creates 3-patterns. n-patterns can be mined after adding analogical columns, `ev_3`,`ev_4`,...,`ev_`n, to `pat_cols` and `temp`.

```
def create_3_pat(pat,good,filename,patterns,events,path):
    ev_cols = ['event','date']
    events_pat = events[ev_cols].loc[events['pseudopatnummer'] \
                == pat]
    events_pat.reset_index(drop=True,inplace=True)
    pat_cols = ['ev_1','ev_2','end']
    patterns_pat = patterns[pat_cols].loc[ \
                patterns['pseudopatnummer'] == pat]
    patterns_pat['pattern'] = patterns_pat['ev_1']+'-' \
                +patterns_pat['ev_2']
    patterns_pat = patterns_pat.loc[ \
                patterns_pat['pattern'].isin(good)]
    for p in patterns_pat.iterrows():
        # keep only events on the same day or later
        events_date = events_pat.loc[ \
                        events_pat['date']>=p[1]['end']]
        if len(events_date.index) != 0:
            # keep only events that do not appear \
                        in pattern p already
            events_date = events_date.loc[ \
                        -events_date['event'].isin( \
                    [p[1]['ev_1'],p[1]['ev_2']])]
            temp = pd.DataFrame(columns = ['pseudopatnummer', \
                        'ev_1','ev_2','ev_3','end'])
            temp['ev_3'] = events_date['event']
            temp['end'] = events_date['date']
            temp['ev_1'] = p[1]['ev_1']
            temp['ev_2'] = p[1]['ev_2']
            temp['pseudopatnummer'] = pat
            temp.to_csv(path+filename,mode='a',header=False, \
                        index=False)
    return None
```

A.3. Threshold optimization

```
def predict_thr(fpr,tpr,thresholds):
    temp = pd.DataFrame([tpr,fpr,thresholds]).T
    temp.columns = ['tpr','fpr','thr']
    ideal = temp[(temp['tpr']>= 0.7) & (temp['fpr'] <= 0.3)]
    if len(ideal.index) == 0:
        ideal = temp[(temp['tpr']>= 0.7) & (temp['fpr'] <= 0.4)]
        if len(ideal.index) == 0:
            return False
        elif len(ideal.index) == 1:
            threshold = ideal['thr']
        else:
            threshold = ideal['thr'].loc[ideal['fpr'].argmin()]
            if isinstance(threshold, float):
                return threshold
            else:
                return threshold[0]
    elif len(ideal.index) == 1:
        threshold = ideal['thr']
    else:
        # find point with derivative closest to 1
        ideal['delta'] = temp['fpr'].loc[ideal.index \
                        + 1].tolist()
        ideal['delta'] = ideal['delta'] - ideal['fpr']
        ideal['2delta'] = temp['fpr'].loc[ideal.index \
                        + 2].tolist()
        ideal['f_delta'] = temp['tpr'].loc[ideal.index \
                        + 1].tolist()
        ideal['f_2delta'] = temp['tpr'].loc[ideal.index \
                        + 2].tolist()
        ideal['deriv'] = np.divide(ideal['f_delta'] \
                        -ideal['tpr'], ideal['delta'])
        ideal['deriv2'] = np.divide(ideal['f_2delta'] \
                        -ideal['tpr'],ideal['2delta'])
        ideal['deriv'].loc[ideal['deriv'] == 0] \
                        = ideal['deriv2'].loc[ideal['deriv'] == 0]
        ideal['diff'] = np.power(1-ideal['deriv'],2)
        threshold = ideal['thr'].loc[ideal['diff'].argmin()]
        if isinstance(threshold, float):
            return threshold
        else:
            return threshold[0]
```