Quantitative and temporal approach to utilising electronic medical records from general practices in mental health prediction

Polchlopek, O.; Koning, N.R.; Buchner, F.L.; Crone, M.R.; Numans, M.E.; Hoogendoorn, M.

Computers in Biology and Medicine 125 (2020) 103973

Contents lists available at ScienceDirect

Computers in Biology and Medicine

journal homepage: www.elsevier.com/locate/compbiomed

Quantitative and temporal approach to utilising electronic medical records from general practices in mental health prediction

Olga Półchłopek [a,*], Nynke R. Koning [b], Frederike L. Büchner [b], Mathilde R. Crone [b], Mattijs E. Numans [b], Mark Hoogendoorn [a]

[a] Department of Mathematics, VU University Amsterdam, De Boelelaan 1081, 1081 HV Amsterdam, The Netherlands
[b] Leiden University Medical Center, Albinusdreef 2, 2333 ZA Leiden, The Netherlands

ARTICLE INFO

Keywords:
Electronic medical records
Temporal pattern mining
Pattern recognition
Mental health classification
General practice

ABSTRACT

This study proposes a framework for mining temporal patterns from Electronic Medical Records. A new scoring scheme based on the Wilson interval is provided to obtain frequent and predictive patterns, as well as to accelerate the mining process by reducing the number of patterns mined. This is combined with a case study using data from general practices in the Netherlands to identify children at risk of suffering from mental disorders. To develop an accurate model, feature engineering methods such as one hot encoding and frequency transformation are proposed, and the pattern selection is tailored to this type of clinical data. Six machine learning models are trained on five age groups, with XGBoost achieving the highest AUC values (0.75–0.79) with sensitivity and specificity above 0.7 and 0.6 respectively. An improvement is demonstrated by the models learning from patterns in addition to non-temporal features.

1. Introduction

Mental health problems (MHP) are relatively common in children and teenagers [1,2], with prevalence rates between 7% and 30% [3–6]. Among teenagers, rates of depression and anxiety have increased by 70% in the past 25 years [7] and the number of university students disclosing a mental illness has grown fivefold in the past decade [8]. Such disorders have a negative impact on everyday life and might lead to repercussions in well-being and functioning, especially if experienced in childhood [9–13]. A substantial number of young patients is recognised in a late state [14–17]. As a consequence, mental health remains insufficiently treated, with a large proportion of children in need not receiving optimal help [18–20].

This is a retrospective with healthcare data from 100 000 children enlisted with general practice centres in the area of Leiden, the Netherlands. It processes Electronic Medical Records (EMR) from general practitioners (GP-s). The dataset consists of coded records (patients, symptoms, consultations, lab results, medication, referrals) and free text mined from the doctors' notes. The data does not involve records of psychological assessments except diagnoses and referrals to specialists. The task is to predict (classify) if patients are going to develop any kind of mental health problem (receive diagnosis, referral or prescription) within a certain time.

Most of the models trained on EMR tackle recognition of cancer [21,22] and cardiovascular diseases [23]. In MHP early models were unsuccessful and progress was made only when using psychosocial data, which is hardly ever gathered for healthy patients, as it is not typically a part of the diagnosing process. In this study some state-of-the-art techniques are used to build models based on non-psychosocial GP records.

EMR allow for retrieving information in a temporal manner, i.e. preserving the chronology of the symptoms raised by the patients. Time series approaches are difficult to apply due to irregularity of visits – a patient can see their doctor twice in one week and then not make another appointment for a year. Recurrent Neural Networks (e.g. long short term memory kind) can preserve the order of events without relying on time as a variable but limited insight can be derived from them; they do not output which combinations of events are predictive. Traditional Machine Learning algorithms, such as linear regression, support vector machines or random forests, are easier to describe in terms of what features impact the outcome the most, but they do not account for the temporal dimension by default.

One way to pass the order of medical history to these models is to create temporal patterns. There are many categories of contributions to the field, focusing on defining and finding frequent patterns in large

* Corresponding author.
E-mail addresses: olga.polchlopek@gmail.com (O. Półchłopek), n.r.koning@lumc.nl (N.R. Koning), f.buechner@lumc.nl (F.L. Büchner), m.r.crone@lumc.nl (M.R. Crone), m.e.numans@lumc.nl (M.E. Numans), m.hoogendoorn@vu.nl (M. Hoogendoorn).

https://doi.org/10.1016/j.compbiomed.2020.103973
Received 17 May 2020; Received in revised form 11 August 2020; Accepted 11 August 2020
Available online 18 August 2020

datasets, then evaluating their performance in tasks such as prediction or clustering [24]. The events can be treated as points in time (assuming no duration) [25] or as time intervals (allowing overlap) [26], and there exist methods of converting one representation to another [27]. The approach presented combines the time-point and the interval representations while retaining the benefits of both. This allows to use all available EMR data, since it contains events with and without duration. It also reverts the typical discovery-then-testing plan by introducing supervision from the classification task already at the mining step. The patterns obtained this way are both frequent and predictive.

The method is based on I. Batal's Minimal Predictive Temporal Patterns [28]. The approach concatenates two or more consecutive and co-occurring events into one that can be used as a binary feature. Sometimes, however, the order and co-occurrence of events cannot be inferred from the data. Therefore, this paper presents a special framework for defining and mining the patterns, designed to account for the complexity of GP derived EMR. There are three main contributions to the approach:

1. Combining time-point and interval representations in a single framework and enrichment technique using expert knowledge of morbidity to infer duration of events
2. Scoring method based on Wilson confidence interval to discover frequent and predictive patterns and limit the number of candidates
3. A framework to mine patterns in a setting where events may overlap, together with a new, vectorised mining algorithm.

Furthermore, this paper aims to target the following questions:

1. Can models trained on general practitioners data give accurate results in a mental health classification task?
2. Can temporal patterns improve traditional Machine Learning methods? What is the quality of the new scoring method?
3. Which Machine Learning method (logit, regression tree, random forest, XGBoost, neural network) provides the best results in a classification task of predicting risk of mental health problems?
4. How far into the future can we predict?

In the course of answering these questions, multiple predictive models were built for mental health problems based on electronic medical records from general practices and new techniques were developed to use this data in the most efficient way. This is a contribution towards developing a decision support system for GP-s to help recognise children at risk of MHP during their visit.

The paper is organised as follows. Section 2 makes a note of the related work, listing similarities and differences. The methodology section provides definitions, together with a mathematical setup, for obtaining features and mining patterns. This framework builds on the approach in [28], yet adapts it to the available data and extends it with a tractable scoring method. Then, the dataset is evaluated and experimental setup proposed in Section 4. The results of the empirical application are presented in Section 5. Section 7 sums up the results and discusses the next steps.

2. Related work

2.1. Models in mental health

In the field of predicting mental health problems early research showed little evidence of success. In 1996 Lewis [29] created a computerised self-assessment system of common mental disorders (PROQSY) which showed short-term improvement in number of mental health consultations compared to lack of such assessment. No difference was reported long term. In Schriger's PRIME-MD system [30] 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 [31] 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 utilised EMR.

Some systems incorporated expert knowledge to create rule-based classifiers, e.g. Masri's and Mah Jani's MeHDES [32]. Such systems required prior knowledge of those rules and were not able to adjust or learn new facts. Other attempts applied decision trees [33], constrained logic programming (if–then clauses) [34], Brain Imagining [35] and Fuzzy Logic [36]. Support Vector Machine, Bayesian Network, Logistic Regression, Radial-Basis Function, Random Forest and Polygenic Scoring were used to infer mood disorders from genome data [37]. A more recent study [38] 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 (AUC[1] of 0.88, 0.9 and 0.78 respectively). Although these are successful studies of predicting mental health, hardly any uses non-psychosocial data.

2.2. Sequential patterns

Allen [39] lists 7 different relations between events that have widely been used [40–44]. 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 [28] and here two of them are relevant: *before* and *co-occurs with*. Such patterns are called sequential because they reflect the chronological sequence of events.

There are many categories of contributions to sequential patterns in EMR, some listed below:

- Frameworks defining patterns from events as points in time [21,25,40] or as intervals [24,26,43,45]
- Methods of transforming time-point series to intervals (temporal abstraction) [24,27,40]
- Data structures [43,46] and mining algorithms for frequent pattern discovery [24,26,43,46]
- Frequent pattern candidate generation and evaluation on a Machine Learning problem [24,26,45] or evaluating patterns while generating them [21,28]
- Human-readable patterns [40]

The advantage of treating events as points in time is that it is easy to recognise sequential relations between them. Interval approaches, however, are better at handling co-occurrences. Most EMR come with a single time reference such as the date of having blood drawn for a laboratory test, and others can be attributed duration, e.g. a cold that lasts for one week. It is therefore important not to focus on one framework but allow to combine both. Temporal abstractions serve to abstract time-point series to symbolic intervals but do not provide solutions to events that do not have duration. The presented approach treats all events as points and duplicates them to reflect duration.

Many mining techniques try to solve the frequent pattern problem, i.e. find the complete set of patterns in a given dataset with respect to a given frequency [46]. They use techniques to remove pattern candidates or templates if they would result in infrequent patterns. This does not guarantee that the patterns mined will be useful for a prediction task, for instance many patients might complain about coughing and

[1] Area Under the [ROC] Curve is interpreted as probability that a classifier will rank a randomly chosen positive instance higher than a randomly chosen negative one. See Fig. 4 for ROC.

then be prescribed cough medicine, making this sequence a frequent pattern, but it would not be expected to be a good mental health predictor. The scoring technique used in this study promotes frequent and predictive patterns and can be used to reduce the number of patterns mined. There are multiple benefits of having human-readable patterns, one of them being discovering knowledge. Here we use interpretability to validate the method with medical experts.

2.3. Minimal predictive temporal patterns of I. Batal

Batal's approach was chosen as the basis of the presented framework because it is designed to mitigate multiple EMR challenges. On top of it, it offers a way of limiting the number of patterns mined, using the target of the prediction task. This ensures that the training data is made of potentially predictive features, which is important when dealing with large and sparse datasets such as the GP derived data. Using frequent patterns instead could introduce noise and would require an additional, iterative step of evaluating the patterns' usefulness after the model is trained, selecting the most useful ones and retraining. This approach is costly in terms of time and computation, so a non-iterative pattern selection process, such as Batal's, is preferred.

A few changes had to be made in order to use this method in the context of GP data and MHP challenges. A lack of a window within which a patient is observed is one of them. It might be desirable to look only at EMR from a fixed interval before diagnosis, as such records are usually more predictive: most diseases, especially those for which decision support is being sought, give first symptoms in a short, defined time. This is not true for mental health disorders. They can emerge rapidly or be caused by a distant childhood experience [3,17,47]. The overlapping of events breaks another assumption of Batal's framework. There are usually multiple health problems that occurred between visits but their start and end dates are not provided (they are recorded under the same date, which is the date of the visit).

The means of reducing the number of patterns introduced by Batal are revisited and a new mining procedure is defined. Such adjustments are necessary, as Batal's assumptions which eased the process – the fixed-width observable time window in medical history and the lack of overlapping events – do not hold.

2.4. Other patterns

Other existing but rejected pattern mining techniques are data stream mining and sequence pattern mining. Data stream mining is suitable for processing sequences of data (streams) such as traffic and sensor data, recognising their drifting and evolution [48]. A stream is expected to be infinite and EMR are not dense enough to be considered as such. Moreover, this technique focuses on most recent patterns and discards older history [49], which is not a preferred approach to predicting mental health. Sequence pattern mining identifies frequent events or sets of events [50] such as buying certain products together [51]. The most frequent patterns can help connect drugs to side effects or symptoms to diagnosing tests but might not be good predictors for MHP. There is more interest in capturing rare but predictive events.

3. Methods

Using patterns in multivariate electronic medical records serves to represent the temporal aspect of the data. Patterns can capture the same information as features (e.g. patient 1 had hydrocortisone prescribed) but also their order in time (e.g. patient 1 suffered from a rash and *then* had hydrocortisone prescribed). Preliminary selection of patterns based on their frequency and expected predictiveness helps to decrease the computational power needed and shorten the execution time.

The method formally detailed below can be distilled to the following. We define a set of EMR *events* (Section 3.1) and for each patient we place them on a timeline (Definition 3). We add *duration* to these events, making them intervals that may overlap (Fig. 1). We look for *n-patterns* — sequences of n events such that each event happened before or together with the next one in the sequence, according to the intervals (Definition 4, Figs. 2 and 3). We calculate *support* of a pattern, which is the number of patients whose EMR contain the pattern, and its *confidence*, which is the share of positive cases among them (Section 3.3). We explain why confidence is an insufficient measure of pattern predictiveness and propose Wilson score transformation, which promotes more common patterns over less common ones (Section 3.4). For each pattern length n, we score $n-1$-patterns, then mine n-patterns. We advise a non-technical reader to focus on the above mentioned parts, although the full framework is given for reproducibility of the approach.

Data engineering methods that help obtain features and events from EMR are listed in Section 3.1. Theoretical framework defining patterns is described in 3.2 and scoring methods are proposed in 3.3. The mining algorithm is detailed in Appendix A. The definitions introduced in these sections are inspired by Batal [28], except Definitions 2, 5, 7 and 10, which are custom-made and constitute the core of the methodology presented in this paper.

3.1. Event selection and feature engineering

An event is the smallest building block of a pattern. What is called an event is strictly dependent on the EMR. If the input data was electrocardiogram values in time, an event could characterise the shape, e.g. the line going up and back down to the same point. GP records consist of laboratory test results, symptoms reported by the patients, diagnoses, prescriptions, referrals and notes made by the doctors. Below are examples of events that can be created from such data.

ICPC codes for symptoms (e.g. `D01` for abdominal pain) and diagnoses (e.g. `R78` for acute bronchitis), free text keywords from notes such as `stress`, `behavioural problems`, `ADHD`, `autism` and `sleeping disorders` and referrals to specialists are all separate events.

Especially sparse data, it is relevant to help the machine treat similar events as such [52] and there is potential to group *medication* according to classification of medication via the ATC framework. The codes are constructed according to the function, type and composition of drugs and the therapeutic subgroup was chosen as a broad enough category, e.g. *hydrocortisone* `D07AB02` and *betamethasone* `D07CC01` fall into `D07****`, which is a group of corticosteroids). An event is having a medicine from a certain code group prescribed. To distinguish the codes, "icpc" or "atc3" are added as prefixes (e.g. `icpc_D01`, `atc3_D07`).

The study is dealing with children's health and the ranges for laboratory tests considered within the normal range depend on age, weight and laboratory used [53]. There exists no easy reference to put the tests into a framework that would enable to compare them between patients. Each laboratory test is therefore conveyed as the name of the test and not the value measured. In order to capture some temporal dimension, events of increased and decreased results (compared to the previous result of the same patient) are used for laboratory tests with numerical values. This is only possible for tests that have been repeated.

In this case it requires less effort to obtain events than features, as most of them can be taken directly from the data. The features are created to compare models trained with and without patterns. Five standard data transformations are applied to create features, one followed by a custom function. These consist of one-hot encoding of categorical variables, calculating counts and frequency (counts per year), measuring the time in between and clustering by hierarchy as explained above.

ICPC, ATC and ATC3 codes are one-hot encoded. This means there is a binary feature for each code, with 1 where the code was found in the EMR and 0 otherwise. Dates of entering and exiting the EMR

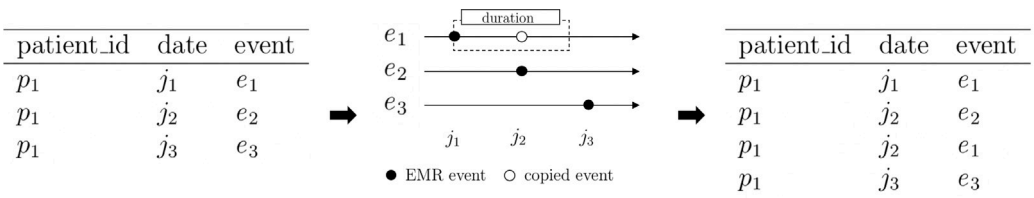


Fig. 1. Duplicating events according to duration.

system serve to calculate the number of GP consultations per year. Each laboratory test is conveyed as the number of times it has been done per patient. *Referrals* to specialists are counted the same way. Additionally, the frequency of having blood drawn (times per year), the minimal time between two tests of a kind and the minimal time between two blood-sampling events are calculated.

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. Function f below converts it to a bounded interval: the inverse provides the bounds and the logarithm distributes values more evenly.

Definition 1. 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 $x = 1$

3.2. Mathematical framework

Let $P = \{1, \dots, n\}$ 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 ith patient and $y_i \in \mathcal{Y} = \{0, 1\}$ is a target class label. We call (x_i, y_i) an *instance* of D. 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. The map ϕ does not have a closed form — it is a composition of all feature engineering transformations. 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 $X_{test}|y_{test}$.

Definition 2. An event E is defined as an instance of an ICPC, ATC usage, referral to a specialist, a laboratory test being performed or its result change (increase/decrease). For every ATC only a corresponding *atc3* group constitutes to an event to reduce the number of patterns mined. Let Σ be a finite set of all permitted events — all ICPC codes, all ATC3 codes, etc.

Definition 3. Let $E_i[t_{start}, t_{end}]$ be an event $E \in \Sigma$ registered for the ith 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 enrichment technique uses the duration of events available in [54], 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). t_{end} is determined based on this research by adding the respective number of days to the date of the visit for non-chronic events. Being diagnosed with any chronic disease is a disturbing event in a child's life and a strong risk factor for mental health problems according to literature [55], hence repeated chronic disease diagnoses were not considered in this step. For other variables (prescriptions, laboratory tests, referrals), $t_{start} = t_{end}$ is assumed.

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

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

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

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

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

The enrichment creates more overlapping events and therefore more patterns to be mined from each patient's EMR. Similarly, more patients will be recognised to have experienced each of the patterns, their number being closer to actual. This will result in patterns being less sparse when added as binary features. The enrichment process and the resulting table are depicted in Fig. 1.

Although for each patient $i \in P$ the series of events $\{E_{ij}\}_j$ can be ordered by non-decreasing date 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) and the true order for such events is unknown.

Definition 4. An event sequence is a series of events where the events are ordered according to their start times:

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

Commonly in sequential patterns mining the events are ordered by start and end time [56] but this method allows for an unknown end time (e.g. for ICPC codes we only know the day of the visit and can estimate the duration) and events without duration (e.g. referrals or laboratory tests). Definition 4 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 7. However, to better understand the mining algorithm described further in Appendix A, it helps to visualise sequences as ordered permutations of sets, as the algorithm operates on vectors rather than single events. The framework is presented below in Definition 5. Here a set consists of events registered within one visit on date j. A permutation of any number of elements of the set is already a sequence (in line with Definition 4). 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 5. Let $\sigma_1, \sigma_2, \dots, \sigma_m$ be permutations and let $[k]$ denote $\{1, \dots, k\}$. The following is an alternative definition of an event sequence:

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

The ith instance can be represented by an event sequence $\mathcal{E}_i$ and represented uniquely by a set $\{\mathcal{E}_i(\sigma) : \sigma = (\sigma_1, \sigma_2, \dots, \sigma_m) \subset (S_{k_{j_1}}, S_{k_{j_2}}, \dots, S_{k_{j_m}})\}$ 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.

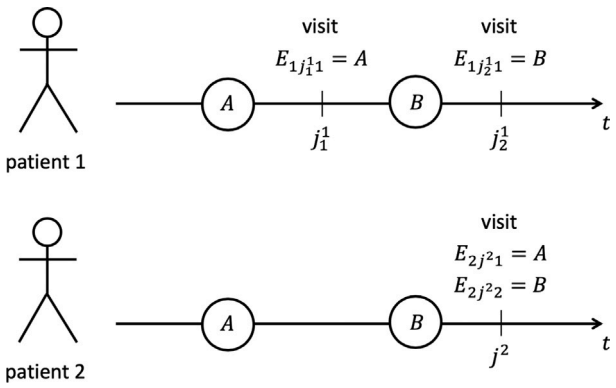


Fig. 2. Timelines with the same events and different distribution of visits.

Definition 6. 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}, in short notation (E_{ijk}, E_{ilm}). Notice how it is allowed for j and l to be the same date. This approach serves mainly to capture situations such as depicted in Fig. 2: both patients observed events A and B but patient (1) reported them on separate visits and patient (2) on one. Patterns (B, A) and (A, B) are created for patient (2) and this allows to pick up the similarity to patient (1). This is also how co-occurring events and overlapping intervals are handled — if patient (2) experienced A and B together, their EMR would stay the same and patterns (B, A) and (A, B) would pick up similarity to yet another patient for whom A and B overlapped. This way the co-occurrence is not explicitly expressed in any pattern but instead captured in presence of 2 patterns. This allows for using the benefits of point-in-time representation of events (simplicity, readability) and retaining their interval character.

Definition 7 expresses this logic in a formal way.

Definition 7. Let E_{ijk} denote the kth event registered for the ith instance on the jth 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 ith 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 $i \in P$. 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 $E_{ipq} \neq E_{ijk} \wedge \cdots \wedge E_{ipq} \neq E_{ilm}$ 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 $\mathcal{P} \subset \mathcal{P}'$.

In other words, if we present the timeline as in Fig. 3, we can select consecutive elements of a pattern by moving up, down or right, but never left (moving left would mean going back in time, breaking the temporal order).

Note that patterns used in a prediction task with training and test sets should only be defined on P_{train}. This applies also to Sections 3.3 and 3.4. P is used for notation simplicity in Definitions 7–11 instead of P_{train}.

3.3. Controlling the number of patterns

Let $\mathfrak{n}$ be the power of Σ. Then, there are $\frac{\mathfrak{n}!}{(\mathfrak{n}-k)!} = \binom{\mathfrak{n}}{k} \cdot k! \geq \mathfrak{n} \cdot k!$ k-patterns that can be created using events in Σ for $k \leq \mathfrak{n}$. The number of patterns witnesses factorial growth with the increase of the length of the patterns and linear with cardinality of Σ. It is computationally ineffective and unscalable to mine all the patterns that may exist.

The algorithm presented in this paper uses the training set to score and select subpatterns for further mining. It assumes that predictive patterns stem from predictive subpatterns. This differentiates it from other pattern mining techniques, which are typically unsupervised and use a predefined template (such as *A before B*) to mine all possible candidates.

Batal introduces some method of scoring, using support as in Definition 8 and confidence defined differently than in 11. However, this is only used for pattern selection and has no role in mining. Instead, a theorem is used which guarantees that at most $n+1$ patterns can be mined from an n-subpattern. This theorem does not hold in GP-derived EMR because of overlapping events, hence a new method and a new algorithm had to be developed.

Definition 8. The support of pattern $\mathcal{P}$ in P is the number of instances in $D|_P = \{(x_i, y_i) \in D : i \in P\}$ that contain $\mathcal{P}$:

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

Definition 9. A temporal pattern $\mathcal{P}$ is called *incoherent* if there does not exist any valid event sequence $\mathcal{E}_i$ for $i \in P$ s.t. $\mathcal{P} \in \mathcal{T}_i$. In other words, $\mathcal{P}$ is incoherent $\Leftrightarrow supp(\mathcal{P}, D|_P) = 0$.

It is irrelevant to mine incoherent patterns. 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 recognised as non-target. Both target and non-target-specific patterns are useful in a classification task. The following definitions describe a method of measuring specificity[2] of patterns.

Definition 10. Let $\mathcal{P}$ be a coherent pattern and let $y_i \in \mathcal{Y}$ for some $i \in P$. 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 11. 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.

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

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

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

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

[2] We use the word "specificity" in Section 3 to refer to the patterns being target-specific or non-target-specific. Outside of this section "specificity" is used alongside "sensitivity" to denote a measure of performance in a classification task (True Negative Rate).

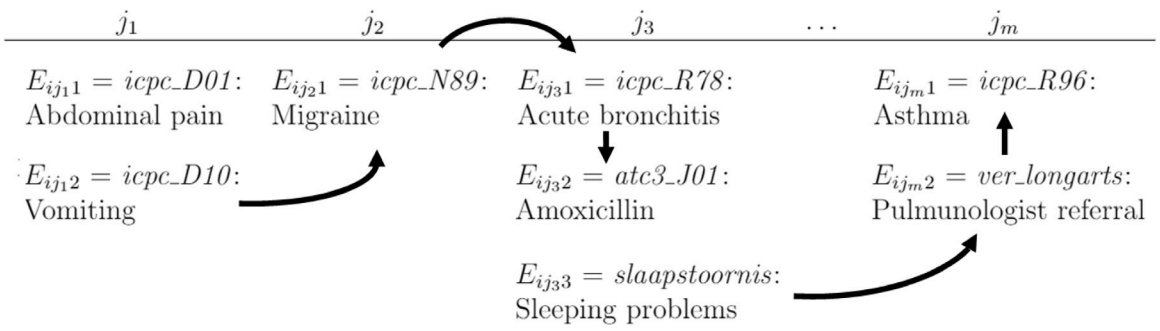


Fig. 3. Examples of a 3-pattern and a 4-pattern on a timeline.

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.

3.4. Wilson score interval

The scoring method presented below tackles the problem of using confidence as a measure of specificity. Batal [28] deals with it by constraining the number of instances that a pattern must cover to be considered. The threshold is called *minimal support* and is chosen locally per class, i.e. the values differ for target and non-target. 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 makes room for improvement. The new method lifts the need for such a constraint by differentiating scores of same-confidence patterns, promoting more common ones.

Wilson score interval is an extension for confidence interval of normal approximation of binomial distribution [57]. Binomial distribution of the target variable can be assumed for the reasons given by [58] but is best illustrated by an example. To know how successful pattern $\mathcal{P}$ is in recognising the target, we count the number of "successes" (target instances) in a subset of patients who exhibit $\mathcal{P}$ just like we would count heads when flipping a coin $supp(\mathcal{P})$ times. Each patient's being a "success" or not is a Bernoulli trial with unknown probability π. Assume $supp(\mathcal{P}) = 20$ and $\pi = 0.4$ which can be thought of as taking 20 patients with $\mathcal{P}$ or flipping a biased coin 20 times.

Wilson interval improves precision of coverage probability of confidence interval. Nominal coverage probability is simply the confidence level set by the definition of the interval, e.g. it is 0.95 for error rate $\alpha = 0.05$. Actual coverage probability is the proportion of the time that the interval contains the true value of interest, e.g. if we repeated the experiment (20 coin flips or checking 20 patients) many (million) times, the true value ($\pi = 0.4$) would fall into the interval 95% of the time and 5% of the time it would not. It is intuitively expected that the 95% confidence interval has 95% coverage probability which is not true. A discrepancy between the actual coverage probability and the nominal coverage probability occurs when approximating a discrete distribution with a continuous one, which is exactly the case here — only 93.61% coverage probability is achieved. The smaller the sample, the higher the difference: taking 10 patients or flipping the coin 10 times yields 89.89%. In the same setup Wilson interval results in 93.63% and 98.16% respectively.

The following definition gives both bounds for the interval but only the lower bound is used.

Definition 12. 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. Examples of how Wilson transformation lowers initial score can be found in Table 1.

Table 1
Examples of Wilson score lower bounds.
Source: Own calculation.

Positive cases	All cases	$conf(\mathcal{P}=1)$	Wilson
78	120	0.650	0.533
5	5	1.000	0.429
4	4	1.000	0.375
10	15	0.667	0.349
2	3	0.667	0.144

This modified scoring serves to choose the best $(n-1)$-patterns for building 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 which features are to be used in the models.

Remark 2. 2-patterns experience a drop in average score compared to 1-patterns, as they are less frequent than single events, and for 3-patterns (and longer) 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 11.

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.

4. Evaluation

The data used in this study was collected for the PIPPI project ("Primary care integrated for identification of psychosocial problems in children" [59]) conducted in the Department of Public Health and Primary Care of Leiden University Medical Centre. It is a collection of EMR from 76 general practices in the Leiden area, gathered, concatenated and preliminarily aggregated by a third party (STIZON[3]). It consists of records of children aged 0–19 from the period 2007–2017 (thus up to and including 31.12.2016).

The collection consists of 7 tables that can be combined using patient identifier and date of the visit:

- *patients*: personal data such as birth year and sex

[3] Stichting Informatievoorziening voor Zorg en Onderzoek.

Table 2
Number of features by category and source table.

Category	Source	Abbrev.	No. of features
ICPC (binary)	Episodes	icpc	927
ATC (binary)	Medication	atc	859
ATC3 (binary)	Medication	atc3	87
Tests (counts and times in between)	Lab results	test	1783
Referrals (counts)	Referrals	ref	100
Free text (binary)	Free text	free	101
No. of visits	Consultations		1
Personal data	Patients		2
Deleted (appeared only in test sets)			−440
Total			3420

Table 3
Number of events by category and source table.
Source: On calculation.

Category	Source	Abbrev.	No. of events
ICPC (codes)	Episodes	icpc	927
ATC3 (codes)	Medication	atc3	87
Tests (codes and change)	Lab results	test	1247
Referrals (codes)	Referrals	ref	100
Free text (text)	Free text	free	101
Total			**2462**

- *episodes*: symptoms and diagnoses coded with International Classification of Primary Care (ICPC) standard (in Dutch) [60]
- *free text*: descriptive symptoms text mined by STIZON from the notes of general practitioners (in Dutch)
- *consultations*: all GP encounters, including phone calls and visits
- *medication*: prescriptions coded with Anatomical Therapeutic Chemical (ATC) standard [61]
- *tests*: any measurements made by the GP or performed in a laboratory
- *referrals*: referrals to specialists (in Dutch)

4.1. Data processing

Some patients were removed from the study due to the missing data (system entry date, gender) or insupportable incoherence (when the calculated *age* or time registered in the system were negative numbers). 27% of the remaining 92 621 patients are positive cases of MHP,[4] which is in line with expectations.

The study is focused on observing patients before being diagnosed and modelling the probability of developing a mental illness in the future. It is therefore important to exclude medical history of target patients within a fixed time window before the first record of MHP and all the post-recognition history of the patient. For each non-target patient a date of stopping the EMR records is defined as either the date they unregistered or the end of data collection for the study.

In all the models, information about the age and sex of the patients is used. Table 2 contains the number of features obtained per category: binary symptoms (ICPC, free text), medication (ATC), counts of laboratory tests and referrals, times between blood tests of the same and different kind, total number of visits per year and personal information (sex and age). Features found only in the test group of patients were removed from training. Analogously Table 3 presents the number of events.

Patients are divided into 5 age groups to reflect stages of children's life and the Dutch education system. For each age group training and testing sets of patients are separated in proportion 7:3.

[4] Identified using all P* code ICPC (psychological), T06* ICPC (anorexia, bulimia), 9 ATC values in N05–N07, 21 referral descriptions in Dutch.

4.2. Research questions and evaluation metrics

To answer if models trained on GP data give accurate results in a mental health classification task, the models are assessed using sensitivity and specificity. Values of sensitivity and specificity above 0.7 are recommended by the American Academy of Pediatrics [62], which means that 70% of healthy and 70% of unhealthy patients should be correctly recognised [59]. Additionally Positive Predictive Value (PPV), Negative Predictive Value (NPV) and F1 score are reported for the best models for clinical relevance.[5] The AUC is compared with benchmark models which are given only the information about sex and age of patients. For qualitative assessment feature importances (or coefficients) are discussed with medical experts.

Comparison between models with and without patterns is required to assess the impact of patterns, therefore two versions of training data is used for the models, one consisting of features (generic models) and one consisting of features and patterns (models with patterns). The patterns are passed to the models as binary features, representing whether a pattern was found in a patient's EMR or not. The compared values are AUC (z-statistic test is conducted [63], as recommended for testing AUC equality [64]) and BIC[6] which considers the number of predictors and sample size. To evaluate the new scoring method, events and patterns with the highest Wilson scores are discussed with clinicians. AUC is used to choose the best Machine Learning model.

Lastly, time windows of 180 and 360 days pre-diagnosis are tried, which means the EMR stopping date is moved by 180 or 360 days backward for each patient. The results in the Section 5 focus on the first run, the other is used for comparison. The time windows are chosen to reflect the actual frequency of GP visits — children and adolescents in the Netherlands visit their doctor at least once a year [65,66].

It would be a good exercise to compare the pattern scoring scheme presented in this paper with Batal's and other known methods. However, it is impossible to use Batal's approach on the same dataset for the same reason that the modified algorithm was proposed — EMR from general practices contain overlapping events and there is no justified fixed time window within which mental health should be assessed, hence the assumptions of the approach do not hold.

4.3. Experimental setup

Patterns are scored within age groups with error rate α for Wilson interval chosen to be 0.01 as recommended in [67]. 2-patterns achieving at least 0.50 and 3-patterns with scores equal to or above 0.55 are

[5] PPV indicates how likely it is that a patient has the condition given that their model outcome is positive, analogously NPV indicates how likely it is that a patient does not have the condition given that their model outcome is negative. PPV is usually much lower than NPV for conditions with low prevalence because of relatively many false positives. Sensitivity and specificity are used for evaluating models on their ability to separate positives and negatives (a good model assigns positives to many actual positives and negatives to many actual negatives) and PPV and NPV are used for understanding individual results (a doctor should know the probability that their patient has the condition). F1 score is a function balancing sensitivity and PPV interpreted as a corrected accuracy when a classic measure of accuracy (% of correct answers) is biased by class imbalance (e.g. for a condition with 1% prevalence a model that assigns negatives to all cases has 99% accuracy). For PPV, NPV and F1 formulas refer to the footnote 8.

[6] Bayes Information Criterion is a measure used to compare models with different sample size or number of parameters (e.g. features). Models with low loss and complexity (number of parameters/samples) are preferred, therefore the loss is penalised by a function of complexity. Here BIC is defined as $m\log(n) + n\log\left(\frac{\mathcal{L}}{n}\right)$ where n is the sample size, m is the number of features and $\mathcal{L}$ is the residual sum of squares (sum of squared differences between binary prediction and true outcome). The model with lower BIC is considered better.

added as target-specific temporal features in age groups 0–3, 4–7, 8–11. The scores differ for different pattern lengths as explained in Remark 2. For the two older groups thresholds of 0.40 and 0.45 are chosen for patterns of length 2 and 3 respectively, as their scores are lower in general. On the other hand, it is impossible to set such thresholds for non-target-specific patterns, as thousands of them have scores of 0. Their number is therefore set to be half of the number of target-specific patterns and patterns with bigger support are prioritised.

Each training set is used to fit six classifiers – logit, SVM, regression tree, random forest, deep neural network and XGBoost – in an attempt to solve a binary classification task. They are chosen to cover wide variety of relations: linear, non-linear, polynomial, rule-based. The models are cross-validated with 3 folds and (where applicable) tuned for AUC using RandomSearch. The following hyperparameters are tried:

- SVM classifiers: linear and RBF kernel, $C \in \{1, 2, 5\}$, $\gamma \in \{0.01, 0.1, 1, \texttt{auto}\}$
- tree classifiers: Gini information criterion and entropy, maximum depth 4, 8 and 12, features per split 30%, 50%, `sqrt` and `None`, minimal samples per split 10, 100 and 500
- random forests: same as trees, number of trees $\in \{10, 50, 100, 1000, 5000\}$, with and without bootstrap
- XGBoost classifiers: same as trees, $\alpha \in \{0, 0.00001, 0.01, 0.1, 1, 10\}$, $\gamma \in \{0, 0.05, 0.1\}$, learning rate $\in \{0.01, 0.03, 0.05\}$, subsample $\in \{50\%, 80\%, 90\%\}$, positive weight = 1 (recommended for 2-class problem), number of estimators decided by running a cross-validated model and settling for the round when the progress stops
- neural networks: activation functions *ELU* and *RELU*, hidden layers $\in \{3, 4, 5\}$, neurons $\in \{40, 50, 100\}$, learning rate $\in \{0.005, 0.01\}$, batch sizes 64 and 128, norm momentum 0.9 and `none`, dropout rate 0.5 and `None`

5. Results

The following sections present the results of pattern mining, accuracy of models with and without patterns and with different time windows, as well as qualitative feature evaluation.

5.1. Generic models

Benchmark models other than trees failed to assign any of the test instances with target class, which means they identified all patients as healthy. This is a usual fault of simple models trained on unbalanced data, i.e. data such that one of the two classes has much more representation (here the ratio of positive instances to negative ones is 1:3).

With few exceptions the generic models performed better than the benchmark, on average making a difference of 0.22 in AUC. Fig. 4 presents ROC[7] curves for all five age groups compared to benchmark. Plotted results are of the best models in each category.

Most of the generic models with time window of 180 days managed to reach sensitivity$\geq$ 0.7 and specificity$\geq$ 0.6, which means that at least 70% of positive and 60% of negative instances were classified correctly. The only models that repeatedly failed to meet the relaxed condition were regression trees. They also achieved the lowest AUC values or close to the lowest. The unquestionable winner in terms of AUC is XGBoost, achieving the desired specificity and sensitivity above 0.7 in two of the groups. Table 4 contains AUC values for models that achieved the highest AUC in tuning per age and algorithm.

[7] A classifier outputs a numerical score for each test case. This score can be used to assign positive or negative outcome given a score threshold (the higher the threshold, the less positive and more negative assignments). The ROC curve is a plot of True Positive Ratio (sensitivity) against False Positive Ratio (1 — specificity) for variable score threshold. For TPR and FPR definitions, see the footnote 8.

Table 4
AUC for generic models.
Source: Own calculation.

Age	Logit	SVM	Tree	RF	DNN	XGB
0–3	0.697	0.714[a]	0.686	0.731[a]	0.696	**0.751**[a]
4–7	0.740[a]	0.729[a]	0.668	0.738[a]	0.756[a]	**0.774**[b]
8–11	0.744[a]	0.736[a]	0.695	0.756[a]	0.759[a]	**0.787**[b]
12–15	0.738[a]	0.702	0.668	0.720[a]	0.737[a]	**0.761**[a]
16+	0.733[a]	0.710[a]	0.685	0.745[a]	0.639	**0.762**[a]

bold = best AUC per age.
[a]Sensitivity above 0.7 and specificity above 0.6.
[b]Sensitivity and specificity above 0.7.

Table 5
Most important features in XGBoost generic model for age 16+.
Source: Own calculation.

Feature	imp.
Number of visits	0.060
Age	0.032
Number of all tests	0.030
Other (ref)	0.023
Antibacterials (atc3 J01)	0.017
Min. time betw. blood draws	0.016
Sex hormones (atc3 G03)	0.016
Levonorgestrel/ethinylestradiol (atc G03AA07)	0.015
Number of unique tests	0.015
Freq. of blood draws	0.013

Feature importance confirms after [68] that the number of visits is a good predictor of mental issues, with the calculated importance at least twice as high as of the other features. In XGBoost models a feature is the more important, the more times it is used to create splits in decision trees. Other aggregating features, namely the number of all laboratory tests, the number of unique tests, the frequency of having blood drawn and the minimal time between those events also appear as highly predictive. Age-specific features selected by XGBoost models are: no illness (icpc A97) or upper respiratory tract illness (icpc R74) for ages 0–3, ear inflammation (icpc H71) for 4–7, medicine for respiratory system (atc3 R06) in the group of 8–11 year-olds, referral to an X-ray in 12-15 and as shown in Table 5, contraceptives for young adults (levonorgestrel/ethinylestradiol, atc G03AA07).

In XGBoost models, random forests and trees each feature can be used in multiple splits to capture non-linear relation, e.g. in one branch of the tree being male could correspond to higher risk and to lower risk in another. It is therefore difficult to infer the direction of the impact the features have, especially when dealing with thousands of trees. Linear models, however, have more interpretable coefficients. Positive value of a significant coefficient characterises a target-specific feature, i.e. if the feature is found (or increases, in case of non-binary features such as the number of visits), the probability of MHP increases. Table 6 consists of signs of some coefficients. The relation of male sex with target variable is mostly positive (except ages 16+), which means that boys should be more prone to registered mental health problems in younger age groups. This dependence was confirmed by other studies, where boys were more often identified with MHP, especially in elementary school [17,69–71].

To summarise, the AUC, sensitivity and specificity values show that non-psychosocial ERM data can be successfully used in MHP classification tasks, with age, the number of GP visits and the number of laboratory tests being the best predictors. Models trained using this data do better than benchmark with statistical significance.

5.2. Models with patterns

The same models were fitted with EMR features and patterns together, meaning that additional, temporal information was added to each of the training instances. XGBoost models have again achieved

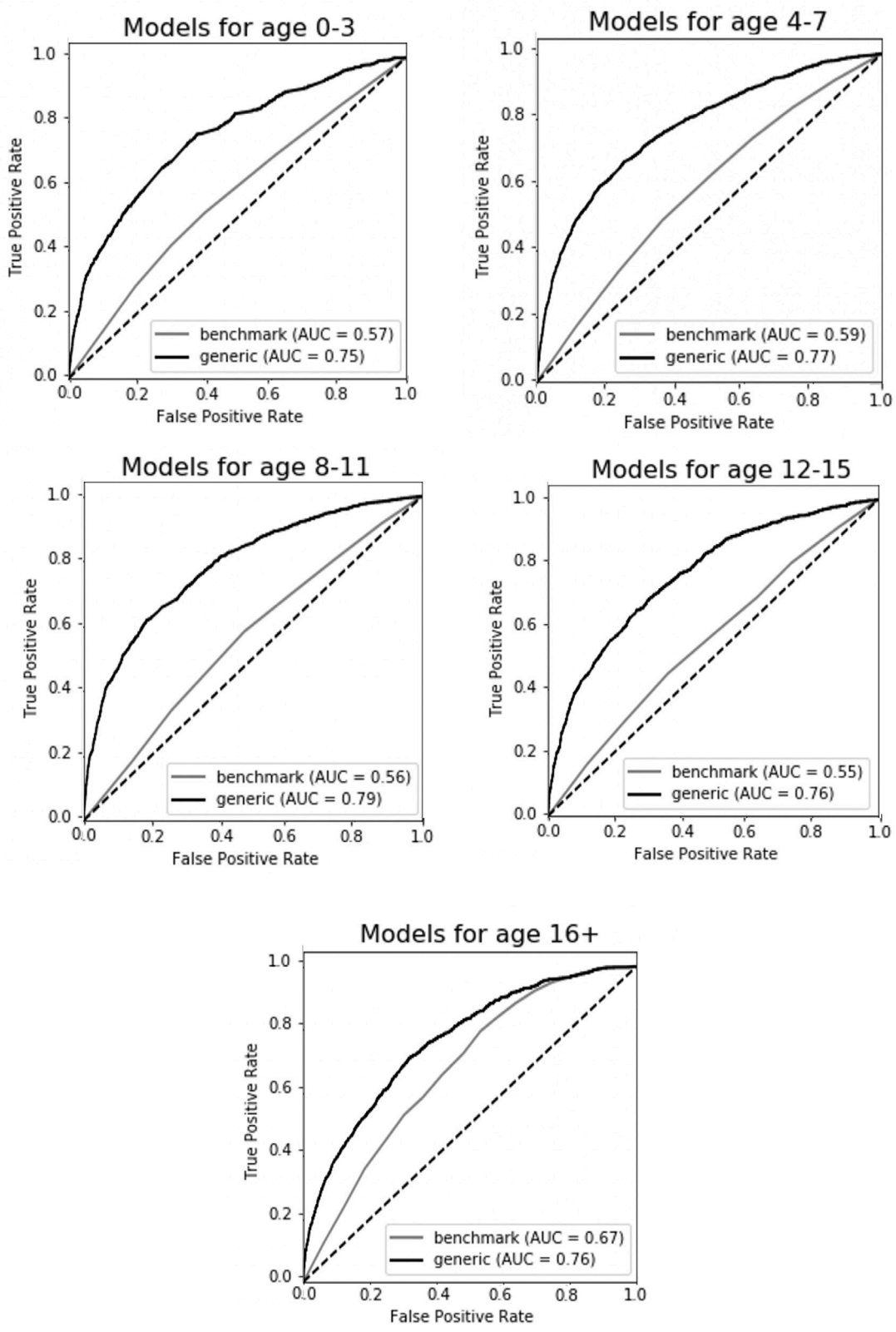


Fig. 4. Receiver operating characteristic curves for generic and benchmark models.

the highest scores within each age group as presented in Table 7. The same two models reached 0.7 specificity and sensitivity.

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. The improvement generated by adding patterns is noticeable but low and the ROC curves almost overlap. Two-tailed statistical z-tests fail to confirm the difference between AUC. BIC values for models with patterns are on average higher than those of generic models. Only three models managed to reach a lower BIC value. Having failed to achieve a better score

Table 6
Signs of selected coefficients in generic logit models.
Source: Own calculation.

Feature	0–3	4–7	8–11	12–15	16+
Age	–	+	–	–	–
Levonorgestrel/ethinylestradiol (atc G03AA07)	n.a.[a]	–	–	+	–
Respiratory med. (atc3 R06)	–	+	+	–	–
Freq. of blood draws	+	+	+	+	–
Min. time betw. blood draws	–	+	+	–	+
No illness (icpc A97)	–	–	–	–	–
Ear inflammation (icpc H71)	+	+	+	+	–
Upper respiratory (icpc R74)	+	–	–	–	–
Sex: male	+	+	+	+	–
Number of all tests	–	–	+	–	–
Number of unique tests	–	–	–	+	+
Other (ref)	–	–	–	–	–
X-ray (ref)	+	–	+	+	+
Number of visits	+	+	–	–	+

[a]There was no record of atc G03AA07 in this age group, hence it was not included in the model.

Table 7
AUC for models with patterns.
Source: Own calculation.

Age	Logit	SVM	Tree	RF	DNN	XGB
0–3	0.699	0.712	0.688	0.736[a]	0.691	**0.754**[a]
4–7	0.740[a]	0.737[a]	0.674	0.740[a]	0.761[a]	**0.777**[b]
8–11	0.744[a]	0.740[a]	0.697	0.763[a]	0.761[a]	**0.788**[b]
12–15	0.738[a]	0.706	0.669	0.727[a]	0.729[a]	**0.759**[a]
16+	0.733[a]	0.711[a]	0.688	0.748[a]	0.640	**0.763**[a]

bold = best AUC per age.
[a]Sensitivity above 0.7 and specificity above 0.6.
[b]Sensitivity and specificity above 0.7.

Table 8
Most important features in XGBoost models with patterns for age 16+.
Source: Own calculation.

Feature	imp.
Number of visits	0.061
Age	0.033
Number of all tests	0.025
Other (ref)	0.024
Sex hormones (atc3 G03)	0.018
Antibacterials (atc3 J01)	0.016
Number of unique tests	0.016
Min. time betw. blood draws	0.015
Freq. of blood draws	0.014
Levonorgestrel/ethinylestradiol (atc G03AA07)	0.013

means that the additional information provided by the patterns was not enough to compensate for increased complexity. Improving the quality of patterns, as well as eliminating correlation by removing subpatterns could be good premises for lowering BIC.

Feature importance in models with patterns (Table 8) to a great extent repeated the outcome of the generic models. Logit models, witnessing slightly lower AUC than XGBoost, rely on different features, as depicted in Table 9. It turns out that patterns played a more important role in linear models, it could therefore be a field worth exploring.

Significant features can be interpreted already in the process of creating patterns thanks to the scoring method. The main aim of this study is to find possible predictors for child MHP, not to draw causal inference. However, some of the predictors found in this study are reported to be related to risk factors for MHP in the literature. For example, increasing and decreasing glucose levels can be found

Table 9
Most target-specific features in logit models with patterns.
Source: Own calculation.

Age	Feature	coef.
0–3	Silver sulfadiazine (atc D06BA01)	1.038
	Constipation med. (atc3 A06)-dermatological antifungals (atc3 D01)-NAN (ref)	0.948
	Chronic tonsil/adenoid infection (icpc R90)	0.937
4–7	Otitis/myringitis (icpc H71)-acute bronchitis (icpc R78)-antipruritics (atc3 D04)	3.174
	Other musculoskeletal injury (icpc L81)-nosebleed (icpc R06)	2.926
	Antipruritics (atc3 D04)-X-ray (ref)-cough and cold (atc3 R05)	2.389
8–11	Chronic tonsil/adenoid infection (icpc R90)	0.514
	Immunisation (icpc A44)	0.478
	Sex: male	0.471
12–15	Acute sinusitis (icpc R75)	0.500
	Chronic tonsil/adenoid infection (icpc R90)	0.425
	Immunisation (icpc A44)	0.406
16+	Tension headache (icpc N02)	0.410
	Paul–Bunnell (test)	0.410
	Cesar therapy (ref)	0.394

Table 10
Most predictive events according to the Wilson score.
Source: Own calculation.

Age	Event	Train	Target	Score
0–3	Gastrointestinal med. (atc3 A03)	120	78	0.532
	Epilepsy (icpc N88)	10	9	0.492
	Campylobacter culture (test)	6	6	0.474
4–7	White blood cells differential (test)	9	9	0.575
	Neutrophils count (test)	9	9	0.575
	Antibiotic and antiviral med. (atc3 J05)	37	29	0.574
8–11	Worms/eggs type I (test)	20	18	0.620
	Worms/eggs type II (test)	20	18	0.620
	Dog allergy (test)	46	36	0.596
12–15	Increasing glucose (test)	12	10	0.462
	Hormone med. (excl. sex hormones) (atc3 H01)	33	21	0.415
	Immunisation (icpc A44)	85	47	0.415
16+	Asthma/COPD (test)	4	4	0.375
	Increasing hepatitis B antigen (test)	4	4	0.375
	Decreasing glucose (test)	60	26	0.283

among the target-specific events in Table 10. They may indicate diabetes, which some studies list as correlated with mental health. For instance, Bădescu [72] claims depression occurrence is two to three times higher in people with diabetes. The aforementioned patterns, "fever (icpc A03)-gastrointestinal medication (icpc A03)" and "fever (icpc A03)-dermatological medication (atc3 D02)-gastrointestinal medication (atc3 A03)" in Table 11, might relate to an increasing awareness of the gut-brain connection and the influence of intestinal microbiota on psychological well-being [73,74].

Examples of events with the highest and lowest Wilson scores can be found in Table 10. Feature names are decoded using ICPC [60], ATC [61] and laboratory test [75] dictionaries, with abbreviations in parentheses to identify their category in line with Table 2.

Table 11 presents frequent 2- and 3-patterns for each group. In some cases the most predictive events from Table 10 form patterns that score even higher, e.g. worms/eggs-worms/eggs (tests for different types) is built from target-specific events for ages 8–11. Longer patterns narrow down the variance of target predictiveness: in group aged 0–3, 32 out of 38 patients who experienced fever before being prescribed drugs for gastrointestinal disorders were identified with mental health problems, and all 14 of those who were also using dermatological medication in between these events constituted to the target.

To summarise, the models with patterns do better than benchmark with statistical significance. Temporal patterns mined using the

Table 11
Most predictive patterns according to the Wilson score.
Source: Own calculation.

Age	Best target-specific			
	Pattern	Train	Target	Score
0-3	Fever (icpc A03)-gastrointestinal med. (atc3 A03)	38	32	0.641
	Fever (icpc A03)-dermatological med. (atc3 D02)- -gastrointestinal med. (atc3 A03)	14	14	0.678
4-7	Antibiotics (atc3 J01)-worms/eggs (test)	13	13	0.661
	Worms/eggs (test)-Giardia lamblia (test)-other (ref)	14	14	0.678
8-11	Worms/eggs (test)-worms/eggs (test)	19	18	0.669
	Nasal med. (atc3 R01)-dog allergy (test)- -cow milk allergy (test)	14	14	0.678
12-15	Urinary tract infection (icpc U35)- -immunisation (icpc A44)	8	8	0.546
	Otitis media (icpc H71)-immunisation (icpc A44)- -hand/foot fracture (icpc L74)	11	11	0.623
16+	Ear concerns (icpc H27)-sex hormones (atc3 G03)	6	6	0.474
	Decreasing glucose (test)-mean cell haemoglobin (test)- -increasing red blood cell count (test)	7	7	0.513

Table 12
Output of the best models with patterns.[8]
Source: Own calculation.

Age	sens.	spec.	PPV	NPV	F1	Confusion matrix		% correct
0–3	0.705	0.683	0.367	0.899	0.483	2442 275	1133 657	68.76
4–7	0.707	0.702	0.527	0.836	0.604	3312 651	1406 1568	70.35
8–11	0.701	0.719	0.600	0.799	0.647	2484 624	972 1460	71.19
12–15	0.701	0.679	0.701	0.708	0.701	899 371	371 871	68.56
16+	0.700	0.678	0.326	0.910	0.445	3992 393	1900 918	68.17

Table 13
Final hyperparameters in XGBoost models with patterns.
Source: Own calculation.

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

proposed approach make a difference in performance of predictive modelling compared to non-temporal models but it is not statistically significant. More data and different models are needed to assess their impact on AUC. Nevertheless, the patterns are human readable and can therefore be qualitatively evaluated, which is a desirable property in EMR models. Some features recognised by the models as important in identifying children at risk of developing psychological problems are events that influence or correlate with mental health according to medical literature, which validates the scoring method.

5.3. The best model

Table 12 consists of accuracy (the percentage of correct predictions on the test set), sensitivity and specificity scores for XGBoost models with patterns, together with other measures calculated from confusion matrices. A confusion matrix informs how many positive and negative values were identified correctly and incorrectly. An ideal classifier would result in numbers on the main diagonal (upper-left corner to lower-right) and zeros otherwise. In majority of models true positives outnumber any kind of mistakes (false positives or false negatives).

XGBoost is a model based on regression trees. A single tree determines which features to use for splits and what values to set the thresholds on, then creates rules that can be read, e.g. "if age ≥ 9 go left, else go right". Final splits end with "leaves" which have assigned scores. The higher the score, the higher the probability of being a target instance. Each tree maps an input instance to one of its leaves. XGBoost ensembles many trees, each to correct the error of the previous tree. Hyperparameters for these models are given in Table 13.

To summarise, of the 6 models tried XGBoost achieves the best results in predicting MHP in young people, reaching AUC above 0.75, sensitivity above 0.7 and specificity above 0.6 (0.7 in some age groups). In juxtaposition with other mental health decision support systems, such as the aforementioned Multilayer Perceptron and Multiclass Classifier from 2017 [38], obtained AUC values are rather low. This is mainly due to the fact that the dataset is based on routine healthcare data from GP-s, in which psychosocial predictors are less thoroughly registered, and is very sparse. Temporal patterns can be used to enhance the performance of the models, both in terms of AUC and predictability. Further feature selection is required to unlock full potential of patterns and reduce correlation.

5.4. Predicting further into the future

In another experiment 360 days of medical history were removed to observe whether the algorithms can predict further into the future. It noticeably improved the performance of the best generic models in terms of AUC, from 0.75–0.79 to 0.78–0.82. Removing a wider time frame caused a drop in the number of distinct EMR events, making the feature matrix less sparse (97% zeros instead of 99%). This might have been the reason for the generic models to improve, as sparse data is not handled very well by the Machine Learning models used.

Models with patterns failed to enhance the improved scores. There are two possible reasons for it. Firstly, patterns add to the sparsity, as they are less common than single events. The drop in performance caused by the introduced sparsity could be higher than the benefit of adding temporal information. Secondly, dropping another 6 months of data could have removed some substantial information, making the created patterns outdated (the last event in the pattern was reported more than a year before diagnosis).

[8] How to read a confusion matrix: TN: true negative (correct), TP: true positive (correct), FN: false negative (mistake), FP: false positive (mistake). TPR (sensitivity) and FPR (1 — specificity) are axes of ROC, TPR and PPV are used to calculate F1.

		Prediction		
		Negative	Positive	
Reality	Negative	TN	FP	RN = FP + TN; FPR = **FP** / RN
	Positive	FN	TP	RP = TP + FN; TPR = **TP** / RP
		PN = TN + FN	PP = **FP** + **TP**	F1 = 2 PPV · TPR / (PPV + TPR)
		NPV = TN / PN	PPV = **TP** / PP	

Table 14
AUC for generic models with constraint on number of features.
Source: Own calculation.

Age	Logit	SVM	Tree	RF	DNN	XGB
0–3	0.696	0.710[a]	0.700	0.737[a]	0.668	**0.749**[a]
4–7	0.738[a]	0.738[a]	0.674	0.747[a]	0.766[b]	**0.784**[b]
8–11	0.738[a]	0.752[a]	0.717	0.773[b]	0.786[b]	**0.812**[b]
12–15	0.709	0.728[a]	0.670	0.740[a]	0.742[a]	**0.788**[b]
16+	0.582	0.627	0.729	0.787[b]	0.675	**0.815**[b]

bold = best AUC per age.
[a]Sensitivity above 0.7 and specificity above 0.6.
[b]Sensitivity and specificity above 0.7.

Table 15
AUC for models with patterns and constraint on number of features.
Source: Own calculation.

Age	Logit	SVM	Tree	RF	DNN	XGB
0–3	0.694	0.710	0.696	0.737[a]	0.694	**0.747**[a]
4–7	0.704	0.735[a]	0.670	0.747[a]	0.771[b]	**0.784**[b]
8–11	0.729[a]	0.737[a]	0.711	0.776[b]	0.782[b]	**0.813**[b]
12–15	0.705	0.722[a]	0.681	0.740[a]	0.730[a]	**0.783**[b]
16+	0.582	0.628	0.733	0.791[b]	0.644	**0.816**[b]

bold = best AUC per age.
[a]Sensitivity above 0.7 and specificity above 0.6.
[b]Sensitivity and specificity above 0.7.

To deal with extra sparsity introduced by the patterns, feature selection can be employed. Stepwise backward elimination based on feature importance was tried in all models. Reducing the number of all features ad patterns to around 2000 made the models perform similarly well, although reaching lower AUC scores than without dimensionality reduction (0.75–0.81 compared to 0.78–0.82). The differences between AUC scores in Tables 14 and 15 are not statistically significant.

To summarise, a long-term MHP identification with GP data (i.e. a 1 year's forecast into the future) does not improve when the patterns are added to non-temporal features. With reduced complexity the two methods perform similarly well.

6. Conclusions for general practitioners

The study has delivered promising results for utilising GP data in identifying children at risk of developing mental health problems. The patterns with high scores were discussed with clinicians and they confirmed having recognised many associations with MHP in their day-to-day work. One way to make use of the models would be to create a software decision supporting tool with user interface and connection to EMR database in a general practice unit. Such a tool could work in the background and assess the score of a patient in real time, during a visit (predicting for one instance is very quick). We recommend that the different age groups are also taken into account in future tools as our results show different features for the different age groups, reflecting different stages in a child's development.

PPV on average above 0.5 and NPV on average above 0.8 in some age groups mean that 50% of cases classified as positive and 80% of cases classified as negative are to be trusted. With this level of uncertainty the tools cannot be used as a diagnostic tool but can serve as a signal for the GP to ask more deeply about the patient's well-being and life, e.g. how it is going at school or at home, and whether the child has friends. An alerted GP might also decide to run psychological questionnaires or refer to a specialist. Or they can take no action at all and monitor the patient and their score across visits — new symptoms will add information to features and a target-specific temporal pattern might be found with the next consultation. It is believed that GP-s have a central role in identification, treatment and referral of mental health problems in children [76], especially because they are able to maintain a long-lasting relationship with their patients [17,65]. Seeing each child's psychosocial development over time on a visual scale could have a beneficial influence on mental issues recognition [18,77].

The methods used in the paper resemble techniques familiar to medical experts. For instance, GP-s use hypothesis (decision) trees to eliminate diagnoses one by one, as well as rule-based assessments with thresholds when reading laboratory test results. A study has shown that experts among medical doctors are more likely to employ pattern recognition in comparison to their non-expert peers [78]. This resemblance, as well as readability of the features can impact adaptation of a new decision support system.

The predictive model built for this study contributed to developing a decision support system for proactive identification of children and adolescents at risk for mental health problems in primary care as part of a project conducted at the Department of Public Health and Primary Care of Leiden University Medical Centre [59].

7. Discussion

Proposed methods of processing EMR resulted in identifying factors predicting mental health, some of them featured in medical literature.

1. The new enrichment technique used expert knowledge to map ICPC codes to their persistence times, which densified the ICPC events. It allowed for mining patterns with higher and more probable supports than if just based on infrequent dates of visits.
2. The new method of scoring based on Wilson interval succeeded at identifying events and patterns that were both frequent and regularly reasonable according to medical experts.
3. The new pattern recognition framework made it possible to use Batal's proven method in the case of overlapping events and the lack of a fixed time window. The new mining algorithm took less than 5 min to run over medical histories of almost 100,000 patients. The algorithm can easily be parallelised to speed up the process even more.

It would be worth 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 the project, which was the reason for developing the modified approach.

Four research questions have been answered in the course of this study.

1. Non-psychosocial GP data proved to be successful in mental health classification.
2. Models trained with temporal patterns accomplished slightly but not significantly better results than generic models in terms of AUC and worse BIC, and the scoring provided insights about potential predictors. Reproducing this study with a different dataset would give more information about the usability of the findings in daily practice.
3. The highest AUC values were achieved by XGBoost, which indicates that the relation of mental health problems and features acquired from EMR is non-linear. Sensitivity and specificity above 0.7 was reached in two age groups.
4. Of the two prediction offsets tried, the method performed better classifying patients in the nearer future (180 days).

Although these are satisfactory results, there are many opportunities for improvement. More expert knowledge should be implemented in data preprocessing: cleaning the EMR, making it uniform and creating a rule-based reference system for the laboratory test results.

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 [28]. Such patterns increase correlation between features but do not enhance performance of the models. This issue has not been addressed in this study.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgement

Funding

This work was supported by ZonMW, the Netherlands, Organization for Health Research and Development (grant 839110012).

Appendix A. Mining algorithm

The mining algorithm starts with a set of events, copies them to represent duration, scores them (events are 1-patterns and can be scored as such) and uses the best ones as the basis for 2-patterns. It then scores 2-patterns and uses the best ones as the basis for 3-patterns, and so on. The pipeline for mining patterns is organised as follows.

1. The events are collected 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.
2. Events for which the duration is known (or can be estimated) are duplicated to account for the duration. Only the dates that already exist in the patient's medical history (the dates of consultations) are used to enrich the event sequence with the lengths of events (duplicating them with other dates would create the same patterns).
3. The events are scored and the most target-specific and non-target-specific of them are chosen. Let us call them *good* for short notation. Each 2-pattern will contain at least 1 *good* event. The idea of restricting the number of subpatterns before mining longer patterns comes from [21].
4. To make 2-patterns, events are sorted by patient and date. They will be picked one by one (in a for-loop) and the names will be added to some of the four lists:
   - all previous events (abbrev. *prev*)
   - all previous "good" (i.e. target-specific) events (abbrev. *prev_good*)
   - previous events on this date (abbrev. *prev_date*)
   - previous "good" events on this date (abbrev. *prev_date_good*)

   If the first event is considered *good*, its name is added to all four lists, otherwise only to *prev* and *prev_date*. The date of the first event is stored as a 1-dimensional variable *date*.
5. Iteration: pick the next **event** name and **date** of the same patient in chronological order. To avoid confusion, elements picked in the current iteration are written in **bold** and stored values from previous iterations are written in *italics*. The following cases may occur:
   a. If **date** is the same as stored *date* (the events were raised during the same visit) and the **event** is *good*, it matches on the right hand side with all the previous (*prev*) events and on the left with all the previous events on *date* (*prev_date*). Then the **event** is appended to *prev_good* and *prev_date_good*.
   b. If the date is the same as *date* but it is not a specific pattern, then it matches analogously with *prev_good* and *prev_date_good*.
   c. If the dates differ, the lists *prev_date* and *prev_date_good* start over. If the **event** is good, its name is added to all four lists, otherwise only to *prev* and *prev_date* as in (4).

   This step is depicted by a flowchart in the figure in Appendix D. Horizontal boxes contain transformations on variables, hexagons mark if clauses, vertical rectangles show how to concatenate vectorised lists to form 2-patterns and *len* is short for vector length. Steps (a) and (b) are also coded in Appendix B.

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. Since the patterns contribute to the models as binary features, spawning identical superpatterns can be significantly reduced by dropping duplicates per patient before starting the algorithm, keeping the earliest ones. This step can be applied no sooner than the 2-patterns have been created.

An important advantage 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. This will scale the runtime linearly by a factor of the number of available cores. The data can be split when mining the patterns, then concatenated for scoring, as in Appendix C. Another speed-improving feature is that the patterns are added as vectors, not one by one.

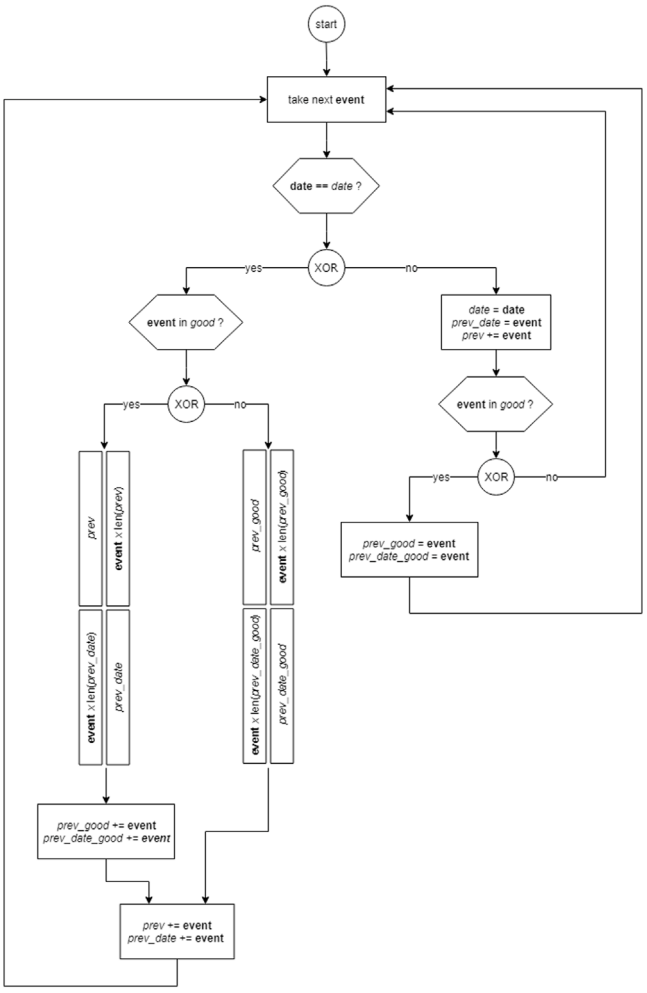
Appendix B. Iterative step of the algorithm

```
if event in good:
    pattern['event_1'] = prev + [event] * len(prev_date)
    pattern['event_2'] = event * len(prev) + prev_date
    pattern['end'] = event_date
    pattern['id'] = patient_id
    # set() used to avoid repetition
    prev_good = list(set(prev_good + [event]))
    prev_date_good = list(set(prev_date_good + [event]))
else:
    pattern['event_1'] = prev_good + [event] \
        * len(prev_date_good)
    pattern['event_2'] = [event] * len(prev_good) \
        + prev_date_good
prev = list(set(prev + [event]))
prev_date = list(set(prev_date + [event]))
```

Appendix C. Splitting the algorithm to run in parallel

```
for n in range(2, N):
    scores[n-1] = score(patterns[n-1])
    good = apply_thresholds(scores[n-1],
        thresholds[n-1])
    for p in patients:  # this can be done in parallel
        patterns[n][p] = make_patterns(patterns[n-1][p],
            patterns[1][p], good)
```

Appendix D. Iterative part of the mining algorithm



References

[1] Kieling, et al., Child and adolescent mental health worldwide: evidence for action, Lancet 278 (9801) (2011) 1515–1525.

[2] Ormel, et al., Mental health in Dutch adolescents: a TRAILS report on prevalence, severity, age of onset, continuity and co-morbidity of DSM disorders, Psychol. Med. (2014) 1–16.

[3] Brugman, et al., Identification and management of psychosocial problems by preventive child health care, Arch. Pediatr. Adolesc. Med. (2001).

[4] Kessler, et al., Lifetime prevalence and age of onset distributions of DSM-IV disorders in the national comorbidity survey replication, Arch. Gen. Psychiatry 62 (6) (2005) 593–602.

[5] Klein, et al., Identification and management of psychosocial problems among toddlers by preventive child health care professionals, Eur. J. Public Health 20 (3) (2010) 332–338.

[6] Reijneveld, et al., Area deprivation and child psychosocial problems - A national cross-sectional study among school-aged children, Soc. Psychiatry Psychiatr. Epidemiol. 40 (1) (2005) 18–23.

[7] Collishaw, et al., Time trends in adolescent mental health, J. Child Psychol. Psychiatry (2008).

[8] C. Thorley, Not by Degrees: Improving Student Mental Health in the UK's Universities, Institute for Public Policy Research, 2017.

[9] A. Goodman, R. Joyce, J.P. Smith, The long shadow cast by childhood physical and mental problems on adult life, Proc. Natl. Acad. Sci. USA (2011).

[10] M.B. Hofstra, J. van der Ende, F.C. Verhulst, Child and adolescent problems predict DSM-IV disorders in adulthood: a 14-year follow-up of a dutch epidemiological sample, J. Amer. Acad. Child Adolesc. Psychiatry 41 (2) (2002) 182–189.

[11] Kim-Cohen, et al., Prior juvenile diagnoses in adults with mental disorder: developmental follow-back of a prospective-longitudinal cohort, Arch. Gen. Psychiatry 60 (7) (2003) 709–717.

[12] 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, J. Child Psychol. Psychiatry 48 (6) (2007) 601–608.

[13] 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, Lancet 380 (9859) (2012) 2197–2223.

[14] E. Garralda, Child and adolescent psychiatry in general practice, Aust. N.Z. J. Psychiatry 35 (3) (2001) 308–314.

[15] Reijneveld, et al., Psychosocial problems among immigrant and nonimmigrant children - Ethnicity plays a role in their occurrence and identification, Eur. Child Adolesc. Psychiatry 14 (3) (2005) 145–152.

[16] K. Sayal, E. Taylor, Detection of child rental health disorders by general practitioners, Br. J. Gen. Pract. 54 (502) (2004) 348–352.

[17] Zwaanswijk, et al., Consultation for and identification of child and adolescent psychological problems in Dutch general practice, Family Pract. (2005).

[18] K. Sayal, Annotation: Pathways to care for children with mental health problems, J. Child Psychol. Psychiatry 47 (7) (2006) 649–659.

[19] N.T. Tick, J. van der Ende, F.C. Verhulst, Ten-year increase in service use in the Dutch population, Eur. Child Adolesc. Psychiatry 17 (6) (2008) 373–380.

[20] Ormel, et al., Mental health in Dutch adolescents: a TRAILS report on prevalence, severity, age of onset, continuity and comorbidity of DSM disorders, Psychol. Med. 45 (2) (2015) 345–360.

[21] M. Hoogendoorn, F. Burkhardt, Machine Learning for the Quantified Self: On the Art of Learning from Sensory Data, Springer, 2018, pp. 54–56.

[22] S.P. Somashekhar, R. Kumarc, A. Rauthan, et al., Abstract S6-07: Double blinded validation study to assess performance of IBM artificial intelligence platform, watson for oncology in comparison with manipal multidisciplinary tumour board – First study of 638 breast cancer cases, Cancer Res. 77 (4) (2017).

[23] S.E. Dilsizian, E.L. Siegel, Artificial intelligence in medicine and cardiac imaging: harnessing big data and advanced computing to provide personalised medical diagnosis and treatment, Curr. Cardiol. Rep. (2014) 16–441.

[24] R. Moskovitch, Y. Shahar, Fast time intervals mining using the transitivity of temporal relations, Knowl. Inf. Syst. 42 (1) (2015) 21–48.

[25] Zhao, et al., Learning from heterogeneous temporal data in electronic health records, J. Biomed. Inform. 1 (65) (2017) 105–119.

[26] D. Patel, W. Hsu, M. Lee, Mining relationships among interval-based events for classification, in: Proceedings of the 2008 ACM SIGMOD International Conference on Management of Data, 2008.

[27] Y. Shahar, A framework for knowledge-based temporal abstraction, Artificial Intelligence 90 (1–2) (1997) 79–133.

[28] Batal, et al., A temporal pattern mining approach for classifying electronic health record data, ACM Trans. Intell. Syst. Technol. (2013).

[29] Lewis, et al., Computerized assessment of common mental disorders in primary care: effect on clinical outcome, Family Pract. 13 (1996) 120–126.

[30] 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, Ann. Emerg. Med. 37 (2001) 132–140.

[31] Rollman, et al., A randomized trial using computerized decision support to improve treatment of major depression in primary care, J. Gen. Intern. Med. 17 (2002) 493–503.

[32] R.Y. Masri, H. Mat Jani, Employing artificial intelligence techniques in Mental Health Diagnostic Expert System, Int. Conf. Comput. Inf. Sci. 1 (2012) 495–499.

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

[34] R.H. Yap, D.M. Clarke, An expert system for psychiatric diagnosis using the DSM-III-R, DSM-IV and ICD-10 classifications, in: Proceedings of the AMIA Annual Fall Symposium, 1996, pp. 229–233.

[35] Kuryati, et al., Investigating machine learning techniques for detection of depression using strctural MRI volumetric features, Int. J. Biosci. Biochem. Bioinform. 3 (2013) 444–448.

[36] Rozita, et al., Employing artificial intelligence techniques in Mental Health Diagnostic Expert System, in: Proceedings of In Computer & Information Science International Conference, 2012 pp. 495–449.

[37] Mehdib, et al., Data mining approaches for genome-wide association of mood disorders, Psychiatr. Genet. 22 (2012) 55–61.

[38] Anujume, et al., Performance analysis of machine learning techniques to Predict Mental Health disorders in Children, Int. J. Innov. Res. Comput. Commun. Eng. 5 (5) (2017).

[39] F. Allen, Towards a general theory of action and time, Artificial Intelligence 23 (1984) 123–154.

[40] F. Hoppner, Knowledge Discovery from Sequential Data, Technical University Braunschweig, 2003.

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

[42] R. Moskovitch, Y. Shahar, Medical temporal-knowledge discovery via temporal abstraction, in: Proceedings of the American Medical Informatics Association (AMIA), 2009.
[43] Papapetrou, et al., Discovering frequent arrangements of temporal intervals, in: Proceedings of the International Conference on Data Mining (ICDM), 2005.
[44] E. Winarko, J.F. Roddick, Armada – an algorithm for discovering richer relative temporal association rules from interval-based data, Data Knowl. Eng. 63 (2007) 76–90.
[45] Moskovitch, et al., Procedure prediction from symbolic Electronic Health Records via time intervals analytics, J. Biomed. Inform. (2017).
[46] Pei, et al., H-mine: Hyper-structure mining of frequent patterns in large databases, in: IEEE International Conference on Data Mining, 2001, pp. 441–448.
[47] Keyes, et al., The burden of loss: unexpected death of a loved one and psychiatric disorders across the life course in a national study, Amer. J. Psychiatry 171 (8) (2014) 864–871.
[48] G. Widmer, M. Kubat, Learning in the presence of concept drift and hidden contexts, Mach. Learn. 23 (1) (1996) 69–101.
[49] G. Hulten, L. Spencer, P. Domingos, Mining time-changing data streams, in: Proceedings of the 7th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Vol. 26–29, 2001, pp. 97–106.
[50] R. Agrawal, T. Imieliński, A. Swami, Mining association rules between sets of items in large databases, ACM Sigmod Rec. 22 (2) (1993) 207–216.
[51] B. Padmanabhan, A. Tuzhilin, Unexpectedness as a measure of interestingness in knowledge discovery, Decis. Support Syst. 27 (3) (1999) 303–318.
[52] s. Trivedi, Z. Pardos, N.T. Heffernan, Clustering Students to Generate an Ensemble to Improve Standard Test Score Predictions, Springer, 2011.
[53] de Vries, et al., Laboratoriumdiagnostiek Bij Kinderen: Een Praktische Handleiding, Prelum Uitgevers, 2015.
[54] Nielen, et al., Estimating morbidity rates based on Routine Electronic Health Records in Primary Care: Observational study, JMIR Med. Inform. 7 (3) (2019) 12.
[55] Bruce, et al., Coping with chronic illness in childhood and adolescence, Annu. Rev. Clin. Psychol. 8 (2012) 455–480.
[56] Papapetrou, et al., Mining frequent arrangements of temporal intervals, Knowl. Inf. Syst. (2009).
[57] E.B. Wilson, Probable inference, the law of succession, and statistical inference, J. Amer. Statist. Assoc. 22 (1927) 209–212.
[58] B. Gerstman, Basic Biostatistics: Statistics for Public Health Practice, Jones & Bartlett Publishers, 2008.
[59] N.R. Koning, F.L. Buchner, M.R. Crone, Primary Care Integrated for Identification of Psychosocial Problems in Children (PIPPI), Vol. 20979, Research Protocol at Leiden University Medical Center, Afdeling Public Health en Eerstelijnsgeneeskunde, 2016.
[60] Waardelijst ICPC-1-2000NL 2011-10-12, 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.
[61] The Anatomical Therapeutic Chemical Classification System with Defined Daily Doses (ATC/DDD), http://www.who.int/classifications/atcddd/en/.
[62] Developmental surveillance and screening of infants and young children, Pediatrics 108 (1) (2001) 192–196.
[63] E. DeLong, D. DeLong, D. Clarke-Pearson, Comparing the areas under two or more correlated receiver operatingcharacteristic curves: a nonparametric approach, Biometrics (1988) 837–845.
[64] X. Zhou, N. Obuchowski, D. McClish, Statistical Methods in Diagnostic Medicine, John Wiley & Sons, Inc., 2002, pp. 180–188.
[65] Wieske, et al., Preventive youth healthcare in 11 European countries: an exploratory analysis, Int. J. Public Health 57 (3) (2012) 637–641.
[66] Bezem, et al., A novel triage approach of child preventive health assessment: an observational study of routine registry data, BMC Health Serv. Res. 14 (1) (2014) 498.
[67] B. Rink, S. Harabagiu, Determining relational similarity using lexical patterns, in: Proceedings of the First Joint Conference on Lexical and Computational Semantics (*SEM), Association for Computational Linguistics, 2012, pp. 413–418.
[68] Richardson, et al., Factors associated with detection and receipt of treatment for youth with Depression and Anxiety Disorders, Acad. Pediatr. 10 (1) (2010) 36–40.
[69] Leaf, et al., Pediatricians' training and identification and management of psychosocial problems, Clin. Pediatr. 43 (4) (2004) 355–365.
[70] J.H. Park, Y.R. Bang, C.K. Kim, Sex and Age differences in Psychiatric Disorders among children and Adolescents: High-risk students study, Psychiatry Inverstigation (2014).
[71] Vogels, et al., Identification of children with psychosocial problems differed between preventive child health care professionals, J. Clin. Epidemiol. 61 (11) (2008).
[72] Bădescu, et al., The association between Diabetes mellitus and Depression, J. Med. Life (2016).
[73] Bercik, et al., The intestinal microbiota affect central Levels of Brain-Derived Neurotropic factor and behavior in Mice, Gastroenterology 141 (1) (2011) 599–609.
[74] E.A. Mayer, Gut feelings: The Emerging Biology of Gut-Brain Communication, Nat. Rev. Neurosci. 12 (8) (2011) 453–466.
[75] SHB NHG Health Base for laboratory results, www.healthbase.nl/media/1310/overzicht-shb-nhg-elementen-v-25-juli_2016-3.xlsx.
[76] Vallance, et al., Managing child and adolescent mental health problems in primary care: taking the leap from knowledge to practice, Prim. Health Care Res. Dev. 12 (4) (2011) 301–309.
[77] Zwaanswijk, et al., Help seeking for emotional and behavioural problems in children and adolescents: a review of recent literature, Eur. Child Adolesc. Psychiatry 12 (4) (2003) 153–161.
[78] Loveday, et al., Pattern recognition as an indicator of Diagnostic Expertise, in: Advances in Intelligent Systems and Computing, Springer, 2013.