sciendo

Int. J. Appl. Math. Comput. Sci., 2021, Vol. 31, No. 1, 107–121
DOI: 10.34768/amcs-2021-0008

amcs

MINING CLINICAL PATHWAYS FOR DAILY INSULIN THERAPY OF DIABETIC CHILDREN

RAFAL DEJA [a], WOJCIECH FROELICH [b,*], GRAZYNA DEJA [c]

[a]Department of Computer Science
WSB University
ul. Cieplaka 1c, 41-300 Dąbrowa Górnicza, Poland
e-mail: rdeja@wsb.edu.pl

[b]Institute of Computer Science
University of Silesia
ul. Będzińska 39, 41-205 Sosnowiec, Poland
e-mail: wojciech.froelich@us.edu.pl

[c]Department of Children's Diabetology
Medical University of Silesia
ul. Medykow 15, 40-752 Katowice, Poland
e-mail: gdeja@sum.edu.pl

We propose a decision support framework (DSF) assisting insulin therapy of diabetic children. Our DSF relies on a medical treatment graph (MTG), which models and graphically represents clinical pathways. Using the MTG, it is possible to plan and adapt medical decisions dependent upon the current health state of a patient and the progress of the treatment. Our MTG fits well with the requirements of clinical practice. The presented work is a cooperative effort of researchers in computer science and medicine. The MTG model has been thoroughly tested and validated using real-world clinical data. The usefulness of the approach has been confirmed by physicians.

Keywords: decision support systems, modeling clinical pathways, diabetes mellitus.

1. Introduction

Diabetes mellitus is one of the most common civilization diseases. Recently, the number of cases has grown rapidly, especially among children suffering from type 1 diabetes. This trend is distressing, as patients with type 1 diabetes must be treated with insulin injections immediately after the diagnosis of the disease. Therapy must be precisely adjusted to the child's energy requirements and lifestyle. Due to the numerous factors influencing the human blood glucose level, according to the American Diabetes Association (ADA), establishing this therapy is difficult (ADA, 2020).

The main challenge in setting up a diabetic therapy is the discovery of care-flow patterns that would be representative enough (Yadav *et al*., 2017). By having those patterns available, it is possible to support the physician in planning diabetic therapy for a particular patient. The problem is known in the literature as mining clinical pathways (CPs).

According to the Australian Queensland Health Board definition, a clinical pathway describes "a standardized, evidence-based medical plan, which identifies the appropriate sequence of clinical interventions, time frames, milestones, and expected outcomes for a homogenous patient group". According to the same organization, the major aim of a clinical pathway is to "support the evidence-based practice, improve clinical processes by reducing risk, and finally, reduce variation in health service process delivery."

The work presented in this paper is in line with one of the most active research areas focusing on mining CPs for chronic care delivery (Zhang and Padman, 2016; Haq

*Corresponding author

et al., 2019; Papiez *et al*., 2019). In Section 2, we give a review of the representative works addressing that problem.

Let us note that, in this paper, we are continuing our previous research on modeling CPs for juvenile diabetic patients. The main limitation of the method we proposed earlier (Froelich *et al*., 2013) is the necessity to define therapeutic templates. Over time, that task alone turned out to be difficult and cumbersome for physicians. Also, our further work proposing the representation of CPs as differential sequences (Deja *et al*., 2015) revealed some limitations. The patients were not initially clustered. This led to a significant number of CPs that were hardly interpretable.

Previously (Deja *et al*., 2017), we focused on mining frequent episodes from temporal data and presented them as CPs. The major limitation of that approach was poor visualization of clinical pathways. The simple graph proposed by Deja *et al*. (2017) was just the direct presentation of frequent episodes. Furthermore, using the frequent episodes approach, it was impossible to filter out less frequent paths or events.

In addition to the above limitations of our previous works, physicians requested to focus our modeling attempts on a single-day therapy, which is a common method used in medical practice (Davidson, 2015). According to medical science, diabetic therapy is based on the so-called "therapeutic day", which is a plan specifying a single day of medical examinations and interventions. After properly setting up that single-day plan, physicians use it repetitively.

To the best of our knowledge, there is no available tool enabling the modeling of the "therapeutic day" at the diabetes onset. At this stage the knowledge about the patient health state, like insulin sensitivity, is limited and therapy has to be established as soon as possible.The absence of such a tool motivates the research presented in this paper. We bridge the gap in the current state-of-the-art by proposing a new approach to modeling the "therapeutic day" of diabetic therapy. Also, we address the issues related to the use of our previous approaches. We meet the requirements stated by physicians asking to make our model transparent and convenient to use in clinical practice.

The modeling approach, which is the contribution of this paper, consists of the following elements:

- a medical treatment graph, which is a model of the "therapeutic day" of a juvenile patient;
- a data mining algorithm enabling the construction of the MTG using raw medical data;
- a set of measures enabling the assessment of diverse clinical pathways represented by the MTG.

Let us note that the application of our MTG brings numerous advantages against competitive approaches. First of all, it provides transparent visualization of alternative medical pathways. Together with the certainty coefficients assigned to the paths of the MTG, it is possible to easily assess the consequences of diverse medical therapies.

The remainder of this paper is organized in the following way. First, in Section 2, we provide a survey of the existing techniques used for the modeling of clinical pathways. The medical problem related to the therapy of diabetic children is described in Section 3. Later, in Section 4, we give the reader a comprehensive presentation of our contribution. Then, based on a real-world case study, we illustrate in Section 5 the work of our approach in practice. In Section 6, we compare our MTG with the other most competitive approaches, i.e., those based on Bayesian networks and Markov decision processes. In Section 7, we validate the MTG using real medical data. Thus, we provide evidence for the credibility of our approach. Section 8 concludes the paper.

2. Decision support systems for diabetic therapy

We position our research in the area of decision support, which is an established field of computer science. In particular, we address the problem of planning sequential actions supporting diabetic therapy (Bennett and Hauser, 2013). To solve that problem, we create an MTG that models the decision process and thus supports physicians in decision making. In the following, we make a review of the existing, alternative decision support systems that serve a similar task.

It is possible to distinguish two types of models of diabetes, namely, non-disease-specific or disease-specific (Bennett and Hauser, 2013). The former models focus on the organizational or economic perspective of a patient's stay in the hospital, e.g., the cost of it, while the latter cover medical therapy, i.e., making medical examinations and administering drugs.

In this study, we consider a clinical, disease-specific model focused on mining clinical pathways from raw medical data. The targeted model is intended to be used by physicians for the diagnosing, controlling, monitoring of the progression, and planing the therapy of diabetes. There are a plethora of diverse types of models serving that purpose. Below, we compare their main characteristics.

Let us first note that the models of diabetes are evaluated qualitatively by physicians during their clinical practice. Therefore, there are no established standards that could be used for the quantitative evaluation of diabetic models (ADA, 2020). However, based on the literature review and the opinions of physicians, we consider the

Table 1. Comparison of models.

Method	Observations	Decisions	Dependencies
mathematical models	variables	variables	mathematical operators
ontologies	terms	terms	terms
fuzzy cognitive maps	fuzzy sets	fuzzy sets	real-valued weights
process mining	events	terms	graphs, operators, weights
templates	events	events	terms
Bayesian network	random variables	random variables	conditional probabilities
Markov decision process	events	terms	probabilities of transitions
MTG	events	events	probabilities of transitions

Table 2. Advantages and limitations of diabetic models.

Method	Reliability	Transparency	Flexibility
mathematical models	excellent	poor	poor
ontologies	poor	good	excellent
fuzzy cognitive maps	poor	excellent	poor
process mining	good	excellent	good
templates	good	good	excellent
Bayesian network	good	good	good
Markov decision process	good	excellent	good
MTG	good	excellent	excellent

following three qualitative criteria that can be used for the comparison of diabetic models.

- Reliability: this criterion assesses whether the model applied represents well the physiological processes governing the glucose–insulin interaction in the human body. High reliability of the model means it has been validated in clinical practice and can be used by physicians for confident planning of insulin therapy. Note, however, that in the case of diabetes there is no perfect, fully reliable model of the disease. This is because of the unique physiological traits of each patient. This means that each model of diabetes is approximate and must be used for the therapy of a particular patient under careful supervision of physicians.
- Transparency: this criterion enables physicians to gain insight into the progression of the disease using the model considered. If the model is transparent, the physician can indicate the reasons that led to the patient's current state and predict the consequences of administering a particular dose of insulin, all without profound mathematical knowledge.
- Flexibility: due to the specificity of human physiological reactions, each employed model of diabetes should be adapted to a particular patient. The flexibility of the model can be achieved by its incremental learning using the data that has been gathered during the initial phase of the given, individual therapy.

Keeping the above criteria in mind, we compare in Table 2 diverse models of diabetes.

Mathematical models rely on formulas expressing the dependencies among diverse variables reflecting physiological processes occurring in the human body. The main one is the glucose–insulin interaction. Mathematical models are considered very reliable. An example model of that type is presented by De Gaetano *et al.* (2008). Although crucial for modeling the progression of diabetes, mathematical models are difficult to interpret by physicians who are usually not familiar with mathematics. The flexibility of mathematical models relies on proper tuning of many parameters. A review of mathematical models of diabetes is given by Palumbo *et al.* (2013).

An alternative to using mathematical models is to ask experts to construct a diabetic ontology (Szwed, 2013). It consists of concepts (nodes of the graph) and relationships (arcs of the graph). Both concepts and arcs are linguistic, medical terms. The main advantage of that approach is the ease with which physicians can reuse knowledge gathered this way. Ontological models might be treated as reliable but quite approximate. That is because of the qualitative terms used for modeling and the so-called semantic gap between ontological terms and the data standing behind them (separation of the given representation scheme from raw data). Even after augmenting the designed ontology by data-driven representations, e.g., fuzzy rules (Szwed, 2013), the obtained hybrid models still suffer from a semantic gap. For that reason, the ontological approach is more suitable to be applied for the construction

of expert-based medical guidelines than for data-driven models of diabetes.

Another approach to the modeling of diabetes is fuzzy cognitive maps (FCMs) (Bourgani *et al*., 2013). In that case, medical events are represented as fuzzy sets. The dependencies among events are modeled as weighted arcs. The real-valued weights measure the rate of a specific causal effect occurring between concepts. When confronted with the clinical practice of diabetic therapy, the cumulative impact of causal concepts on an effect concept used by FCMs turned out to be unsuitable for our purpose. The issue is that the model does not properly represent the relationship between the measurements of glycemia and the following insulin injections. Also, the iterative approach to the reasoning does not represent well the actual temporal dependencies among medical events occurring in diabetic therapy.

Another approach to modeling diabetes relies on extracting information from process logs. The technique is called process mining. Using that approach it is possible to discover models relying on Petri nets (Weijters *et al*., 2006). The issue is that the obtained model might be heavily obscured by incidental, less representative events. This limits the flexibility of the model. Process mining was used to construct causal nets (Augusto *et al*., 2016). The proposed approach extracts useful information from the hospitalization database gathered during medical therapy. On that basis, a graphical model of clinical pathways was constructed. The limited possibility of modeling complex relationships between patient states underlying diabetic therapy is, from our point of view, the main limitation of that approach. Another limitation of process mining is the assumption that the training data considered do not contain noise (Weijters *et al*., 2006). A formal specification along with all the necessary assumptions for using the process mining technique is presented by Huang *et al*. (2012).

The approach based on Bayesian networks (BNs) enables the probabilistic modeling of diabetes (Marini *et al*., 2015). The BN approach assigns conditional probability tables to the graph nodes, which are random variables related to medical observations and decisions. BNs are a very efficient tool for modeling uncertainties embedded within CPs. Let us, however, note that the interpretation of BNs might not be easy for physicians. That is due to the necessity of interpreting conditional probability tables assigned to the nodes of the network. In Section 6 we make an in-depth comparison of the BN approach with our MTG.

Another approach represents diabetic therapy in the form of Markov models (Elghazel *et al*., 2007; Yang *et al*., 2012; Bennett and Hauser, 2013; Zhang *et al*., 2015; Mattila *et al*., 2016). In particular, the Markov decision process (MDP) can be efficiently used to determine an optimal therapy (policy) (Schaefer *et al*., 2005). Similarly as for the BN, the approach requires gathering a large amount of data (Huang *et al*., 2012). In addition, Markov models are hard to be learned incrementally (Elghazel *et al*., 2007). This means that the probabilities of transitions between an MDP's states have to be recalculated using all available data, also those that have arrived recently.

Another limitation of the MDP is the Markov assumption that the state of the model at time t depends only on the information available at time $t-1$. In Section 6 we compare our approach with Bayesian network and the Markov decision methods.

It is also worth noting that the models of diabetic therapy can be constructed using multi-criteria optimization methods. The goal, in that case, is the optimization of treatment and care protocols taking into account non-disease-specific criteria like the cost of treatment and others. For example, the optimization of medical templates using an evolutionary algorithm was proposed by Funkner *et al*. (2017). A minimax optimization model was developed to generate optimal input parameters for the developed model of CPs (Ozcan *et al*., 2011). Also in that case, the proposed approach was designed with the intention of optimizing the non-disease-specific aspects of health care. Recently, a mixed-integer linear programming-based approach for day-level scheduling of CPs has been proposed (Schwarz *et al*., 2019). The approach used a multi-criteria objective function considering several hospital-related aspects; however, also in that case, the proposed method targeted mainly the optimization of health care management.

Let us also note that some works propose grouping patients' data aiming at improving the quality of the obtained models. The approach proposed by Zhang *et al*. (2015) is similar to ours; however, it clusters patients' sequences (temporal data). The first difference of our approach with respect to the work of Zhang *et al*. (2015) is that in our study we cluster patients into cohorts using static data describing patient clinical state at the submission. Using this type of clustering, we obtain reduction of patients' diversity within cohorts. In addition, Zhang *et al*. (2015) present only the most probable pathway to physicians. Using our MTG, it is possible to observe deviations from the most probable pathway the patient can potentially follow during medical treatment.

A review of works devoted to modeling diabetes is given by Bennett and Hauser (2013) or Aspland *et al*. (2019). A comprehensive comparison of diverse probabilistic models can be found in the work of Barber (2012)

Table 3. Static variables.

Feature	Medical meaning
Age	Age of the patient at the onset
Sex	0 (female) or 1 (male)
Weight	Patient's weight at onset
C-peptide	Insulin secretion
CRP	Certificate of inflammation
PH	ACID based balance

3. Medical context of the computational problem

As mentioned in Introduction, the problem we address in this paper is supporting physicians in planning effective insulin therapy at the onset. The objective of this therapy is stabilization of the patient's blood glucose level (BGL) within an acceptable range, which is called normoglycemia. The targeted stabilization should be accomplished as soon as possible. This is the reason why identification of the proper therapeutic procedure becomes a challenge, both from medical and computational points of view.

Let us first note that each diabetic clinical therapy, independent of the patient considered, relies on a series of insulin doses that should lead to keeping the BGL within a normal range. The adjustment of these doses is the issue that physicians face in clinical practice.

To assess the effectiveness of insulin injections administered by physicians, the patient's BGL is measured several times a day. In this way, insulin doses and glycemia measurements mold a sequence of medical events that, in theory, should lead to long-term normoglycemia.

The initial, first insulin dose the physician administers is based on the patient's energy requirements (a number and content of meals) and the patient's clinical state evaluated upon admission to the hospital. Also, some personal data about the patient are taken into account. Those are the patient's weight, age, and some other data presented in Table 2. As those do not change over time, following the medical literature (Marini *et al.*, 2015), we relate them to static variables characterizing each of the patients considered in Table 3.

For this research, according to medical standards, we define the notions of hypo-, hyper- and normoglycemia (ADA, 2020). The ranges of the BGL characterizing each of those notions are presented in Table 4. Note that they are dependent on the pre- or post-meal period the BGL was measured. We assigned numerical values to the related medical terms. The mapping is presented in Table 5.

Another issue the physician faces is related to the standardization of insulin doses. The so-called pre-meal insulin ratio is calculated as delivered insulin units per 100 kcal of the meal (a balanced diabetic diet is used in the hospital). Moreover, the insulin pre-meal ratio should be related to the patient's weight (specifically 100 kg of the weight). This way, the pre-meal insulin ratio is calculated using 100 kcal of the meal and 100 kg of the body weight. The obtained value is rounded. For example, when considering the before-breakfast period, the patient (see Table 6) got 3.5 units of insulin per 240 kcal (i.e., 1.46 per 100 kcal). Since the patient's weight was 23 kg, the insulin ratio was calculated as 6.3 and rounded to 6 units (based on 100 kg of weight and 100 kcal of the meal).

In Table 6 we give an example sequence of medical events gathered for an anonymous patient. By I and G we denoted (dynamic) variables related to insulin injections and glycemia measurements, respectively.

4. Mining clinical pathways

The approach we propose in this paper consists of five major stages.

1. First, we group patients into representative cohorts (clusters). As explained in Section 3, we use static variables for that purpose, i.e., those that do not change their values over time. We follow here the medical conviction that patients from the same cohort are treated similarly, which is in rapport with the common clinical practice (Zhang and Padman, 2016).

2. We define the notions of an event and a clinical sequence of events. According to those definitions, for each of the previously obtained clusters, we prepare a set of clinical sequences.

3. We define the MTG as a graphical model generalizing clinical sequences. At this stage, we also calculate the values of specific measures related to the MTG.

4. For each of the patients' cohorts, we train a separate MTG using the previously gathered data. For that purpose, we provide a dedicated algorithm.

5. Finally, we use the trained MTG as a decision support tool assisting the physician while planning diabetic therapy of new patients.

In the following, we proceed to a detailed explanation of our approach.

4.1. Grouping patients. Before grouping patients into cohorts, the values of the static variables must be normalized. This way, the influence of each variable on the clustering process becomes the same. To accomplish that, we use a simple min-max normalization that turned out to be well-suited for the problem we address (García

Table 4. Glycemic ranges and their clinical meaning.

Glycemia [mg/dl]	Clinical meaning before breakfast	before other meals	after meal
< 70	hypoglycemia	hypoglycemia	hypoglycemia
$[70, 90]$	normoglycemia	normoglycemia	normoglycemia
$(90, 100]$	mild-hyperglycemia	normoglycemia	normoglycemia
$(100, 140)$	mild-hyperglycemia	mild-hyperglycemia	normoglycemia
$[140, 200]$	mild-hyperglycemia	mild-hyperglycemia	mild-hyperglycemia
> 200	hyperglycemia	hyperglycemia	hyperglycemia

et al., 2015). The parameters, i.e., the minimum and maximum values of each variable, are provided by physicians.

After the normalization, we cluster the static data using the fuzzy c-means method proposed by Dunn (1973). For that purpose, we use the Euclidean distance between data instances. The main advantage of using fuzzy c-mean clustering is that the method calculates for each data instance the degree to which it belongs to each cluster. This means that, for each of the patients considered, we obtain a vector of values which are the degrees to which the patient belongs to the distinct clusters. This vector is provided to physicians, who approve the assignment of a patient to one of the cohorts.

4.2. Clinical events and sequences. Let us define a medical event $u \in U$ as a pair $u = \langle V_i = v, \tau \rangle$, where $V_i \in V$ is a variable and v denotes the value that V_i takes on at time τ. In other words, we say that an event u occurs at time τ when the variable V_i obtains a certain value v from its domain $\mathrm{dom}(V_i)$ at a particular time τ. The set U is the universe of all possible events.

At this stage of research, we assume $V = \{G, I\}$, i.e., we consider only those variables related to the measurements of the BGL (variable G) and insulin ratios (variable I). The domain of G contains the discretized values of the BGL provided in Table 5, i.e., $\mathrm{dom}(G) = \{1, 2, 3, 4\}$. The domain of I is the set of positive integer values determined by the insulin ratio described previously in Section 3.

Let us define a clinical sequence as $s = \langle u_{\tau_1}, u_{\tau_2}, \ldots, u_{\tau_n} \rangle$, were τ_i is the real-time at which an event occurs. The length of s depends on the period the patient stays in the hospital. By S we denote the set of all those sequences. The clinical sequences defined in the aforementioned way serve as the source data for the training of the MTG.

The next step of our modeling approach is related to the time flow. Note that the patient's state highly depends on the time of meals. That, in turn, depends on a particular patient. For that reason, as suggested by Hripcsak *et al*. (2015), we decided to sequence time. However, in our study, we do not sequence the entire period of therapy.

Table 5. Blood glucose level discretization.

Blood glucose level	Discrete value
Hypoglycemia	1
Normoglycemia	2
Mild-hyperglycemia	3
Hyperglycemia	4

According to medical knowledge, we sequence time as shown in Table 7, within a single therapeutic day of a patient. This is in accordance with the discrete time scale used by physicians for the planning of daily insulin therapy.

As presented in Table 7, the patient's therapeutic day is partitioned concerning the predefined time intervals which are related to the meals eaten by the patient. Note that the time intervals may overlap, which is in accordance with clinical practice. This way, instead of dealing with the continuous-time flow, physicians plan therapy according to a specific, discrete time scale.

To sequence time, we create the set of labels $T = \{t_1, t_2, \ldots, t_{11}\}$ and map them to the consecutive time intervals provided by physicians. Table 7 illustrates the created mapping. Thus, we construct a discrete time scale with the time horizon limited to a single therapeutic day. Furthermore, to map medical events to the discrete time scale, we define a function $t : \mathrm{RT} \to T$, where RT denotes the domain of real-time. This means that each u_τ that occurs in real-time τ is mapped to a new, discrete time scale as $u_{t(\tau)}$.

As shown in Table 7, the events related to glucose measurements and insulin injections may occur solely at certain periods. In particular, all insulin injections occur at meal periods, all glycemia measurements, in turn, may be labeled only by 'before meal' or 'after meal' terms. Note also that by introducing the discrete time scale, we abstract not only from the continuous-time flow but also from the particular day at which a medical event occurs.

4.3. Medical treatment graph. Let us define the MTG as a directed acyclic graph $\mathrm{MTG} = (N, E, \sigma, \omega)$, where N is the set of nodes, $E \subseteq N \times N$ is the set of

Table 6. Example of raw clinical data.

Time	Description	Value
7:55	G: glycemia measurement	139 mg/dl
8:00	I: insulin injection	3.5 units
8:05	Breakfast	240 kcal
10:55	G: glycemia measurement	189 mg/dl
11:00	I: insulin injection	2 units
11:05	Second breakfast	170 kcal
13:55	G: glycemia measurement	65 mg/dl
14:00	I: insulin injection	4 units
14:05	Lunch	380 kcal
16:55	G: glycemia measurement	71 mg/dl
17:00	I: insulin injection	4.5 units
17:05	Dinner	480 kcal
19:55	G: glycemia measurement	109 mg/dl
20:00	I: insulin injection	3 units
20:05	Supper	190 kcal
22:00	G: glycemia measurement	66 mg/dl

edges representing pairwise node-to-node dependencies. Functions $\sigma : N \to [0,1]$ and $\omega : E \to [0,1]$ assign real-valued weights to each node and edge, respectively. The semantics of the MTG are explained below.

Let us consider $U_t \subset U$ as a subset of events that occur at time $t \in T$ (in the discrete time scale related to the patient's therapeutic day). We distinguish within U_t the subset of those events $N_{tj} \subset U_t$ determined by a particular variable and its value. This means all $u \in N_{tj}$ refer to the same variable G or I, assuming a certain constant value of glucose or insulin, respectively.

We assume $N_{tj} \in N$ is the node of the MTG, where the index t refers to the time period of the daily therapy and the index j refers to the unique pair of the given variable and its value. This means that the set N_{tj} contains similar events, i.e., those that occur at the same period of the therapeutic day and in addition refer to the same variable and value.

Let

$$\sigma(N_{tj}) = \frac{\text{card}(N_{tj})}{\text{card}(U_t)}$$

estimate the probability of an event from N_{tj} in the group of events from U_t. The value of $\sigma(N_{tj})$ plays the role of the weight of the node N_{tj} within the MTG. Let us consider now the mutual dependencies between nodes. Let us define the edge of the MTG as an ordered pair $E_{tjk} = \langle N_{tj}, N_{(t+1)k} \rangle$, where N_{tj}, $N_{(t+1)k}$ are the nodes related to the sets of events occurring at time t and $t+1$, respectively. Note that, for the sake of clarity, in the case of edges, we use time as a superscript.

Let $S_t \subset S$ be the set of the shortest possible subsequences consisting of only two consecutive events u_t, u_{t+1}. Let us distinguish from S_t those sequences S'_t that match the given pair of neighboring nodes of the MTG. We define that set as $S'_t = \{u_t, u_{t+1}\} | u_t \in N_{tj}, u_{t+1} \in N_{(t+1)k}$.

Table 7. Periods of daily therapy.

T	Description	Period	Event
t_1	before breakfast	[6:00–10:00]	G
t_2	breakfast	[6:00–10:00]	I
t_3	after breakfast	[9:00–12:00]	G
t_4	second breakfast	[9:00–12:00]	I
t_5	after second breakfast	[11:00–15:00]	G
t_6	lunch	[11:00–15:00]	I
t_7	after lunch	[14:00–17:00]	G
t_8	dinner	[14:00–17:00]	I
t_9	after dinner	[16:00–20:00]	G
t_{10}	supper	[16:00–20:00]	I
t_{11}	after supper	[19:00–23:00]	G

Let

$$\omega(E_{tjk}) = \frac{\text{card}(S'_t)}{\text{card}(S_t)}$$

estimate the probability of the consecutive events occurring within the clinical sequences. The function assigns the weights of the edges E_{tjk} of the MTG.

To extend the interpretation of alternative pathways within the MTG, we introduce a specific certainty coefficient. For an edge of the MTG, we define

$$\text{cer}(E_{tjk}) = \frac{\omega(E_{tjk})}{\sigma(N_{tj})}.$$

Note that the certainty coefficient describes the distribution of events along the edges starting at the given node.

Let us assume $p = [p_1, p_2, \ldots, p_n]$ is any path within the MTG, where p_i is the node selected from the MTG and $p_i \in N$, $1 < n \leq 11$. In this case, the index i points the place of the node within the path. Note that p is a clinical pathway that conforms to the definitions provided in Section 1.

We scale up ω aiming at the evaluation of any pathway within the MTG, i.e.,

$$\begin{aligned}\omega(p) &= \sigma(p_1) \cdot \prod_{i=1}^{n-1} \frac{\omega\langle p_i, p_{i+1}\rangle}{\sigma(p_i)} \\ &= \sigma(p_1) \cdot \text{cer}(p_1) \cdot \ldots \cdot \text{cer}(p_n) \\ &= \sigma(p_1) \cdot \text{cer}([p_1, \ldots, p_n]).\end{aligned}$$

We scale up also the certainty coefficient for the pathway of any length as $\text{cer}(p) = \prod_{i=1}^{n-1} \text{cer}(p_i, p_{i+1})$.

By using functions $\omega(p)$ and $\text{cer}(p)$, physicians can assess the credibility of any pathway within the MTG. They can also filter from the MTG those paths less likely to occur, i.e., those related to the exceptional medical cases. Assuming $\omega_{\min}$ is a threshold given by

physicians, it is possible to produce a sub-graph MTG' $= (N', E', \sigma, \omega)$ for which $\omega(p) > \omega_{\min}$ for any p.

To verify this idea, we performed experiments generating different MTGs for different values of ω_{min}. The resulting MTGs were provided to physicians, who selected the most useful one for further application. This way, it was possible to adjust the most suitable value of $\omega_{\min}$ for each of the clusters considered.

Note also that it is possible to transform the MTG to a single pathway, representing the most likely course of diabetic therapy. For that pathway, we have $p_{\max} = \arg\max_{p \in P} \omega(p)$.

4.4. Constructing the MTG. To construct the MTG from data, we propose Algorithm 1. We assume that the initial MTG is empty. This means that the content of MTG concepts is gathered on the fly.

The algorithm searches through the list of sequences of medical events. Every event in a sequence is the candidate for a node in the graph, and each pair of events is a candidate for an edge of the graph. They will become a node and an edge if they are not already registered in the graph.

First, in Lines 2 and 3, the algorithm initiates the collections N and E, which are used for storing nodes and edges of the MTG, respectively.

Later on, the algorithm iterates through the clinical sequences (the loop starts in Line 4) and events within them (the loop starts in Line 5). The clinical sequences are given as an input to the algorithm in the form of the array S. The first index of that array, denoted by j, refers to the sequence considered, and the second one, denoted as i, indicates the i-th event within the j-th sequence and refers to time τ. The sequencing of time occurs in Line 6.

Then, the algorithm searches through the collections N and E, checking whether they contain a particular event (Line 12) and edge (Line 16) detected in the j-th sequence. That loop starts in Line 11.

If the node or edge is found within the MTG, the algorithm increments the related counters NCount and ECount (Lines 14 and 18). Otherwise, the node or the edge is added to the corresponding collections (Lines 22 and 26).

Finally, in Lines 31–36, the algorithm iterates through the constructed MTG to calculate 'cer' and ω.

Let us note that the algorithm has a linear computational complexity concerning both the number of the patient's sequences and the number of events within the sequence.

Algorithm 1. Constructing the MTG.

Require: S—set of clinical sequences, w—number of clinical sequences

1: Function GraphBuild(S, w)
2: N = null; NCount $\leftarrow 0$; {a collection of nodes}
3: E = null; ECount $\leftarrow 0$; {a collection of edges}
4: $l = 1$;
5: **for** $j = 1$ **to** w **do** {for each sequence}
6: **for** $i = 1$ **to** length($S[j]$) **do** {for each event}
7: $l = l + (i \mod \text{card}(T))$;{determine the offset}
8: $t = t(\tau_i)$;
9: node $= S[l][t]$; {create a node}
10: edge $= \langle \text{node}, S[l][t+1] \rangle$; {create an edge}
11: N_{exists} = false; {lacking node}
12: E_{exists} = false; {lacking edge}
13: **for** $k = 1$ **to** l **do** {for the added nodes}
14: **if** $N[k][t] ==$ node **then**
15: {Is the node added?}
16: NCount$[k][t]$++; N_{exists} = true; {a number of nodes}
17: **end if**
18: **if** $E[k][t] ==$ edge **then**
19: {Is the edge added?}
20: ECount$[k][t]$++; E_{exists} = true; {a number of edges}
21: **end if**
22: **end for**
23: **if** not N_{exists} **then**
24: $N[l][t]$ = node; NCount$[l][t] = 1$;
25: {adding the node}
26: **end if**
27: **if** not E_{exists} **then**
28: $E[l][t]$ = edge; ECount$[l][t] = 1$;
29: {adding the edge}
30: **end if**
31: **end for**
32: **end for**
33: **for** $j = 1$ **to** l **do**
34: **for** $i = 1$ **to** card$(T) - 1$ **do**
35: $\sigma[j][i] = \text{NCount}[j][i] / \sum_{k=1}^{l} (\text{NCount}[k][i])$
36: $\omega[j][i] = \text{ECount}[j][i] / \sum_{k=1}^{l} (\text{ECount}[k][i])$
37: cer$[j][i] = \text{ECount}[j][i]/\text{NCount}[j][i]$
38: **end for**
39: **end for**
40:
41: **return** MTG

4.5. Applying the MTG in clinical practice. Below, we provide an instruction facilitating the use of our MTG in clinical practice.

1. After admission to the hospital, the patient should be assigned to one of the cohorts considered. Let us remember here our assumption that the

Table 8. Clustering validity check.

No. of clusters	3	4	5	6	7	8	9
CH index	45.97	45.30	45.72	45.48	52.83	50.21	49.97
XB index	0.76	0.65	0.81	0.72	0.62	0.68	0.90

Table 9. Cluster centroids (values after denormalization).

#	Patients	Weight	Age	Sex	C-peptide	CRP	PH	Within-cluster variation
1	16	56.3	15.4	0	0	0.02	0	0.44
2	6	45.7	13.9	1	1	0.01	0	0.18
3	23	24.4	6.7	0	0	0.01	0	0.28
4	27	29.7	8.7	1	0	0.01	0	0.24
5	8	48.7	12.6	0	1	0.02	0	0.39
6	15	40.4	11.1	0	0	0.02	1	0.29
7	7	25.2	7.6	1	0	0.01	1	0.18

patient belongs, to a certain degree, to each of the constructed cohorts, as explained in Section 4.1. For that purpose, the related uncertainty degrees calculated by the fuzzy c-means method are shown to physicians. Based on that and the medical expertise, the physician assesses which cluster is the most representative for a given patient and makes an ultimate assignment.

2. Now the MTG relevant to the patient's cluster is presented to the physician. Depending on the actual state of the patient, the physician may filter from the MTG those paths less likely to occur, forming in this way a sub-graph. The filtering is usually performed several times, allowing the physician to analyze alternative pathways.

3. The physician interprets the obtained MTGs and on that basis constructs the plan of therapy.

4. As therapy proceeds, the physician confronts the current state of the patient with the related part of the MTG. On that basis, the physician adjusts the diabetic therapy.

5. Case study

Let us first note that, due to the confidentiality of the personal information conveyed by medical data, hospitals are not allowed to make them publicly available. This is especially valid in the case of diabetic children. For that reason, we were restricted to, for validation purposes, the data of 102 patients gathered at a single hospital—the Diabetes Center located in Katowice, Poland.

The statistical properties of the static data are shown in Table 10. Let us note that despite a single child much older and heavier than the others, we did not detect outliers in data. For this kind of data, centroid-based clustering is usually a good choice.

Table 10. Data statistics.

	Min	Max	Mean	Stdev
Weight	10.0	85.9	36.06	16.96
Age	1.1	17.7	9.81	4.34
Sex	0	1	0.40	0.49
C-peptide	0	1	0.16	0.37
CRP	0	1	0.05	0.22
PH	0	1	0.23	0.42

In accordance with the presented approach, the available data were normalized and partitioned using the fuzzy c-means clustering algorithm. The number of clusters was chosen using the Calinski–Harabasz (CH) criteria (Calinski and Harabasz, 1974). Also, a Xie–Bieni index was considered (Xie and Beni, 1991). The highest value of the CH index (and the lowest of XB) was reached for 7 clusters, and so this number of clusters was chosen for our purposes (see Table 8 for details). The cluster centroids and within-cluster variation we obtained are presented in Table 9. The stability of the clustering has been verified by changing the random initialization over several runs. We repeated the clustering 10 times with different random initialization of the cluster centers. The Rand index, calculated as the number of pairs of patients distributed in the same clusters or always in different clusters to the total number of pairs, was 0.97. The stability (and quality) of the clustering was satisfactory.

For each cluster, the historical data related to the BGL and pre-meal insulin dosages were discretized and converted into clinical sequences, as explained in Section 4.2. Then we used the proposed Algorithm 1 to construct MTGs.

Due to space limitations, we present in Fig. 1 only a sub-graph for the first cluster. The cohort here can be characterized as mainly female, completely insulin-dependent, without diabetic ketoacidosis, older,

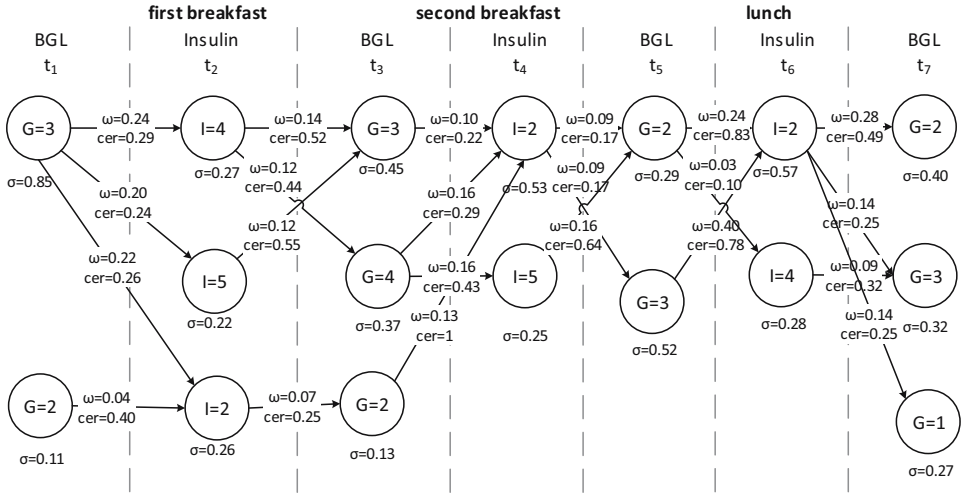


Fig. 1. Example of a medical treatment graph.

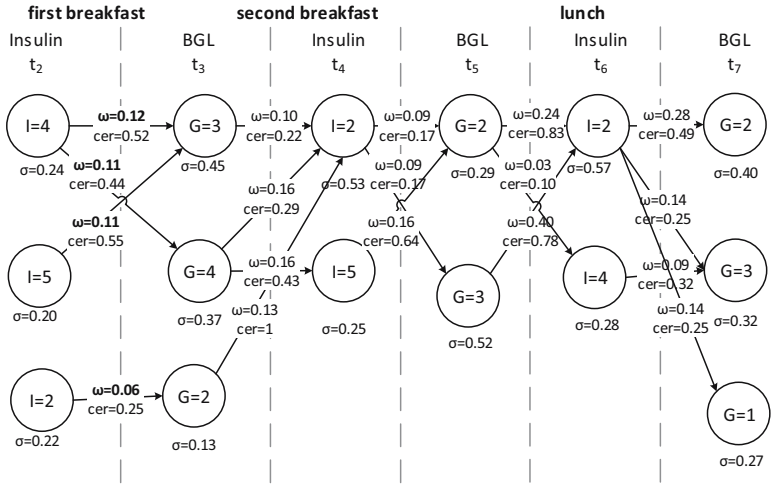


Fig. 2. MTG updated for $\langle G = 3, t_1 \rangle$.

and heavier than the others.

The pathways were filtered out using the threshold $\omega_{\min} = 0.00015$. As explained previously, the value of that parameter was suggested by physicians after performing several trials. For the sake of clarity, we simplified in Fig. 1 the notation, which is self-explanatory.

Following the first path in the graph, we interpret it in the following way. A group of 85% of patients from the first cluster elevated mild hyperglycemia in the morning $\langle G = 3, t_1 \rangle$, whereas 29% of them within the first breakfast were administered around 4 units of insulin per 100 kcal per 100 kg of body weight. On the other hand, 40% of patients with normoglycemia in the morning $\langle G = 2, t_1 \rangle$ got 2 pre-meal insulin units for the first breakfast. Around 3 hours after the first breakfast $\langle G = 4, t_3 \rangle$, an excess of the BGL was observed in around 37% of the patients, and 32% of them were administered 4 units of insulin before $\langle I = 4, t_2 \rangle$.

Some conclusions drawn from the graph have an obvious medical explanation. When starting with normoglycemia in the morning, a lower insulin dose is required for the first breakfast and it is easier to keep the proper BGL after the meal (see events in t_1, t_2, t_3). It can be noted that insulin doses vary, especially during the first breakfast, and this is mainly because of the different BGL before meals. Also, the body response differs even

after applying the same insulin dose $\langle I = 2, t_6 \rangle$. It is also worth noting that glycemia is usually above normal before and after the first breakfast.

In the next stage of our experiments, we considered only the pathways exhibiting the highest value of the ω coefficient. It should be noted (Table 11) that the daily treatment path usually starts with mild-hyperglycemia but at the end of the day it decreases to the normal level. Therefore, the insulin dose is much higher in the morning than in the evening. Depending on the cluster, we observe changes in insulin doses during the day. For example, for the second cluster, the BGL remains approximately normal the whole day.

In the last column of Table 11, the values of ω calculated for selected pathways are quite small. This was, however, expected by physicians because of the numerous fluctuations of the BGL that usually occur during diabetic therapy.

As the physician proceeds with the therapy of a particular patient, our MTG can be shortened (cut off) using the currently recognized patient state. More precisely, let us consider the node N_{tj} and the set of nodes $N_{(t+1)*}$ connected with it by the set of edges E_{tj*}. Assuming that the event represented by the node N_{tj} already occurred, the graph can be shortened to present only the consecutive paths, i.e., coming out from N_{tj}. After shortening the graph, the σ coefficients for consecutive nodes were recalculated as $\sigma(N_{(t+1)k}) = \omega(E_{tjk})$. Consequently, ω of each edge $E_{(t+1)jk}$ had to be adjusted proportionally to the σ distribution. The values of the 'cer' coefficient obviously remain unchanged.

To give an example, after the mild-hyperglycemia observed before the first breakfast $\langle G = 3, t_1 \rangle$, the MTG was accordingly updated. That part of the MTG is shown in Fig. 2. Please note that the values of σ and ω in the figure are rounded to the hundredth fractional part, and the cer coefficient has been calculated before rounding (and is the same as before shortening). Thanks to the performed update, the MTG was simplified, enabling physicians to focus their attention on therapy following the event that already occurred.

6. Comparative study

In this section, we analyze differences among the three most competitive approaches to modeling medical pathways, namely, our MTG, Bayesian networks (BNs), and Markov decision process (MDPs). For that purpose, we used data gathered for the first cluster of our patients. For the sake of readability of the comparison, we produced all three models for the first three time steps t_1, t_2, t_3 of the therapeutic day corresponding to the first breakfast period, namely a before breakfast BGL, before-breakfast insulin injection and after-breakfast BGL.

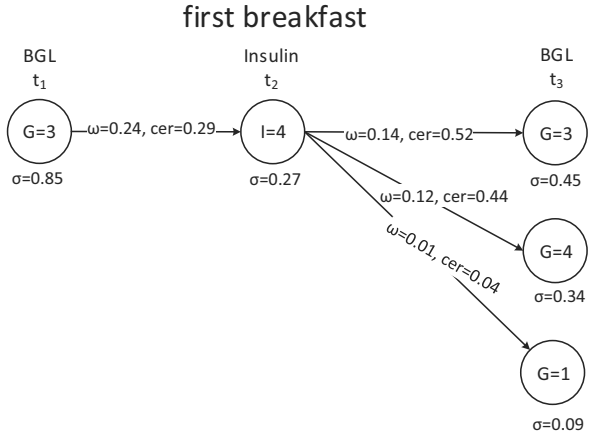
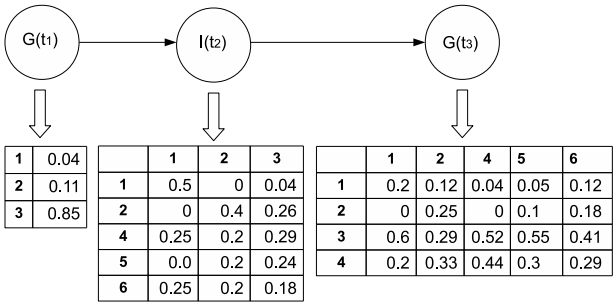


Fig. 3. Example MTG.



1	0.04
2	0.11
3	0.85

	1	2	3
1	0.5	0	0.04
2	0	0.4	0.26
4	0.25	0.2	0.29
5	0.0	0.2	0.24
6	0.25	0.2	0.18

	1	2	4	5	6
1	0.2	0.12	0.04	0.05	0.12
2	0	0.25	0	0.1	0.18
3	0.6	0.29	0.52	0.55	0.41
4	0.2	0.33	0.44	0.3	0.29

Fig. 4. Example Bayesian network.

The produced MTG is presented in Fig. 3. For the sake of clarity, we consider at the time t_1 only a single node that refers to the most probable amount of BGL measured at that time. Also, for t_2, we depicted only a single, most promising consecutive node related to the injection of four units of insulin. Later, at time t_3, we consider all possible consecutive nodes referring to the alternative values of G. The σ coefficient assigned to nodes gives the physician explicit information on the probability of the related event. In turn, the values ω and 'cer' enable us to evaluate the likelihood of transitions between events that occurred within the therapies of similar (with respect to their static data) patients. It becomes clear, by looking at the MTG, that physicians are able to identify not only a single pathway best supported by data, but also other pathways, alternative in terms of their probabilities of occurrence.

An alternative to using the MTG is the application of the Bayesian network, presented in Fig. 4. In this case, variables G and I are assigned to the nodes of the graph, so it is not possible to differentiate events as nodes of the graph. The probability distribution tables corresponding to nodes are depicted below them. As can be noted, the Bayesian network contains similar information as the MTG. The main difference between both approaches lies

Table 11. Expected pathways (# is the cluster number).

#	Pathway	$\omega(p)$
1	$G=3I=4G=3I=2G=2I=2G=2I=2G=1I=1G=2$	0.0015
2	$G=2I=2G=2I=2G=2I=2G=2I=2G=1I=1G=2$	0.0018
3	$G=3I=6G=3I=2G=2I=2G=1I=1G=1I=1G=2$	0.0002
4	$G=3I=6G=3I=6G=3I=6G=1I=1G=1I=1G=2$	0.0001
5	$G=3I=6G=3I=6G=3I=6G=3I=6G=1I=1G=2$	0.0011
6	$G=3I=6G=4I=6G=4I=6G=3I=5G=1I=1G=2$	0.0001
7	$G=3I=6G=2I=4G=2I=2G=2I=2G=1I=1G=2$	0.0005

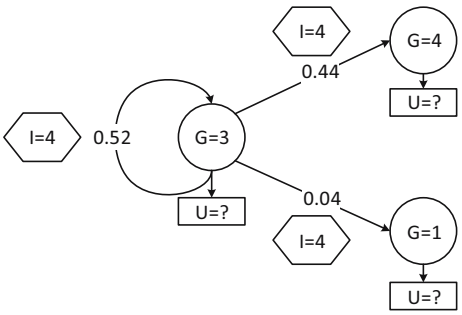


Fig. 5. Example Markov decision process.

in better transparency of our MTG, which can be easily interpreted by physicians. The distribution of events is directly visible in MTG. Furthermore, interpreting paths of the MTG as the clinical pathways allows physicians to easily adjust the current therapy as its different alternatives are clearly shown within the MTG.

In Fig. 5 we show a model of the Markov decision process representing the discussed part of the therapeutic day. As can be noted, the MDP deals with state nodes and decision nodes that relate to the BGL and insulin injections, respectively. The edges of the MDP model are marked by the probabilities of the related state-to-state transitions. Note also that the MDP contains a loop, i.e., it is not an acyclic graph as the MTG and the BN are. The MDP aims at finding decisions that maximize the expectation of some accumulative reward (utility). Therefore, the MDP requires defining a utility function that cannot be defined in the case of diabetic therapy. Since the patient's state is evaluated subjectively by physicians considering a number of diverse medical factors, we are not able to calculate the rewards required to be given for the MDP. Using a distance between normoglycemia and the current BGL could be considered a simplified proxy for utility. However, this would not fully reflect the long-term oriented deviation of the BGL. For the above reasons, the unknown values of the utility function are denoted in Fig. 5 by a question mark.

Table 12. Mean value of κ for 5 learn and test trials.

Cluster	1	2	3	4	5	6	7
κ	5.1	3.8	4.1	3.3	3.3	4.3	3.0

7. Validation

As mentioned in the literature review, there is no established measure that could be used for quantitative evaluation and comparison of different models of medical pathways.

However, specifically, for the validation of our MTG, we designed a benchmarking procedure based on the cross-validation technique. We randomly partitioned all available 102 clinical sequences into the training set containing 80% of them and the testing set containing the rest of them. For the training set, we produced 7 clusters (as was previously chosen) and the corresponding MTGs. To eliminate the noise (exceptional medical situations) involved in data, we filtered the obtained MTGs using $\omega_{\min} = 0.000006$. This parameter was thoroughly adjusted in cooperation with physicians.

For validation purposes, we define a therapy matching coefficient

$$\kappa = \text{avg}_{s' \in S^c} \text{length}(s'),$$

where S^c is the set of patients' sequences from the testing group assigned to cluster c. The higher value of κ indicates that a longer clinical sequence matches any pathway within the MTG.

For the patients from the testing set, we calculated κ. The results of the experiments performed for each of the clusters are presented in Table 12. We underline that the obtained results relate only to the longest continuous clinical sequences. From that perspective, the 3–5 steps ahead of medical therapy supported by the MTG can be interpreted as a good result.

Finally, we asked our physician for qualitative evaluation of the proposed approach. The initial classification of a new anonymous patient into one of the existing clusters was straightforward. Then, the MTG (see Fig. 1) of the first cluster was presented to the physician. As the starting point of therapy, the physician proposed

a daily insulin dose and the doses of pre-meal insulin. Concerning the example patient, the physician estimated, on the basis of the MTG that the daily insulin ratio is ca 10 units per 100 kcal of a meal and 100 kg of patient weight. The physician initially chose a higher dose, taking into account mild-hyperglycemia in the morning and the possibility to adjust the dose later.

Note that insulin requirements often decrease during therapy, and the patient clinical state is changing over time. The patient finally finished the day with the following clinical sequence: $G = 3I = 4G = 4I = 2G = 2I = 2G = 3I = 2G = 3I = 1G = 1$. Hyperglycemia after the first breakfast was observed despite a relatively high rate of the insulin dose and hypoglycemia in the evening (despite a relatively low rate of insulin dose). According to the MTG, normoglycemia after the first breakfast occurred along normoglycemia in the morning and a relatively low insulin ratio. At this stage of therapy, the physician decided not to change the treatment and insulin dose distribution. The obtained MTG reveals that the decrease in the BGL after the first breakfast can be achieved without increasing the insulin dose (see Fig. 1). Also, decreasing the already very low insulin dose before the supper is not recommended. Later on, the following sequence was observed: $G = 3I = 4G = 3I = 2G = 1I = 2G = 1I = 2G = 2I = 1G = 2$, and the glucose balance was improved. It means that the MTG helped the physician to decide on keeping insulin doses unchanged.

The next day, because of the observed normoglycemia in the morning, the physician decided to reduce the insulin dose for the first breakfast, as the MTG suggested. The patient, however, finished the day with the following sequence: $G = 2I = 2G = 4I = 2G = 2I = 2G = 2I = 2G = 1I = 1G = 1$ (so again with hyperglycemia after the first breakfast). After two subsequent days of the therapy, the patient ended up with the following sequence: $G = 2I = 4G = 3I = 2G = 1I = 2G = 2I = 2G = 2I = 2G = 2$, which was only partially covered by the MTG. Therefore, the MTG was helpful only to some degree, namely, in those parts that matched the occurred sequence.

The major conclusions coming from the above validation are the following:

- The proposed approach supports the initial classification of patients to appropriate groups. This information helps to compare the patient's state with the other patients, and thus makes the planning of the patient's initial therapy substantially easier.
- The MTG allows physicians to follow and adapt medical decisions for each insulin application.
- The physician found the possibility of visualizing the consequences of therapy changes, e.g., of reducing the insulin dose for a given meal, very useful. Also, the distribution of insulin doses over the therapeutic day can be adjusted easier when using the MTG.

8. Conclusions

In this paper, we proposed a new approach to modeling CPs of diabetic therapy. Our method proposes abstracting from raw medical data at diverse levels. First, it introduces a symbolic time scale aiming at the representation of the typical therapeutic day. Second, our approach generalizes groups of similar medical events as the nodes of the proposed medical treatment graph. Finally, by counting events that co-occur, the proposed method creates the edges of the MTG. Later on, those edges can be filtered, enabling further abstraction from the noise involved in data. By the proposed abstractions, we developed our MTG as a powerful tool used by physicians in their clinical practice. Let us also mention some limitations of our approach. Firstly our MTG concerns only the pathways related to daily medical treatment. The adaptation to night therapy, during which the patients consume no food, requires further investigation. We also must admit that, due to its data-driven nature, our method can be deemed less reliable than the mathematical models known from the literature. We consider two possible directions for further research. The first one is the modeling of pathways using more data, especially those that can be retrieved from the continuous glucose monitoring system. Modeling pathways leading to extreme hypo-and hyperglycemia is another problem we would like to address.

References

ADA (2020). Children and adolescents: Standards of medical care in diabetes—2020, *Diabetes Care* **43**(Suppl 1): S163–S182.

Aspland, E., Gartner, D. and Harper, P. (2019). Clinical pathway modelling: A literature review, *Health Systems* **0**(0): 1–23.

Augusto, V., Xie, X., Prodel, M., Jouaneton, B. and Lamarsalle, L. (2016). Evaluation of discovered clinical pathways using process mining and joint agent-based discrete-event simulation, *Proceedings of the 2016 Winter Simulation Conference, Arlington, USA*, pp. 2135–2146.

Barber, D. (2012). *Bayesian Reasoning and Machine Learning*, Cambridge University Press, Cambridge.

Bennett, C.C. and Hauser, K.K. (2013). Artificial intelligence framework for simulating clinical decision-making: A Markov decision process approach, *CoRR* **abs/1301.2158**.

Bourgani, E., Stylios, C., Georgopoulos, V. and Manis, G. (2013). A study on fuzzy cognitive map structures for medical decision support systems, *in* M. Nikravesh *et al.* (Eds), *Forging New Frontiers: Fuzzy Pioneers II*, Springer, Berlin/Heidelberg, pp. 151–174.

Calinski, T. and Harabasz, J. (1974). A dendrite method for cluster analysis, *Communications in Statistics—Theory and Methods* **3**(1): 1–27.

Davidson, M. (2015). Insulin therapy: A personal approach, *Clinical Diabetes: A publication of the American Diabetes Association* **33**(3): 123–135.

De Gaetano, A., Hardy, T., Beck, B., Raddad, E., Palumbo, P., Bue-Valleskey, J. and Pørksen, N. (2008). Mathematical models of diabetes progression, *American Journal of Physiology:. Endocrinology and Metabolism* **295**(6): E1462–79.

Deja, R., Froelich, W. and Deja, G. (2015). Differential sequential patterns supporting insulin therapy of new-onset type 1 diabetes, *Biomedical Engineering Online* **14**(1): 13.

Deja, R., Froelich, W., Deja, G. and Wakulicz-Deja, A. (2017). Hybrid approach to the generation of medical guidelines for insulin therapy for children, *Information Sciences* **384**(C): 157–173.

Dunn, J.C. (1973). A fuzzy relative of the ISODATA process and its use in detecting compact well-separated clusters, *Journal of Cybernetics* **3**(3): 32–57.

Elghazel, H., Deslandres, V., Kallel, K. and Dussauchoy, A. (2007). Clinical pathway analysis using graph-based approach and Markov models, *ICDIM 2007 Proceedings, Lyon, France*, pp. 279–284.

Froelich, W., Deja, R. and Deja, G. (2013). Mining therapeutic patterns from clinical data for juvenile diabetes, *Fundamenta Informaticae* **127**(1): 513–528.

Funkner, A.A., Yakovlev, A.N. and Kovalchuk, S.V. (2017). Towards evolutionary discovery of typical clinical pathways in electronic health records, *Procedia Computer Science* **119**: 234–244.

García, S., Luengo, J. and Herrera, F. (2015). *Data Preprocessing in Data Mining*, Intelligent Systems Reference Library, Vol. 72, Springer, Cham.

Haq, A., Wilk, S. and Abelló, A. (2019). Fusion of clinical data: A case study to predict the type of treatment of bone fractures, *International Journal of Applied Mathematics and Computer Science* **29**(1): 51–67, DOI: 10.2478/amcs-2019-0004.

Hripcsak, G., Albers, D. and Perotte, A. (2015). Parameterizing time in electronic health record studies, *Journal of the American Medical Informatics Association* **22**(4): 794–804.

Huang, Z., Lu, X. and Duan, H. (2012). On mining clinical pathway patterns from medical behaviors, *Artificial Intelligence in Medicine* **56**(1): 35–50.

Marini, S., Trifoglio, E., Barbarini, N., Sambo, F., Di Camillo, B., Malovini, A., Manfrini, M., Cobelli, C. and Bellazzi, R. (2015). A dynamic Bayesian network model for long-term simulation of clinical complications in type 1 diabetes, *Journal of Biomedical Informatics* **57**: 369–376.

Mattila, R., Siika, A., Roy, J. and Wahlberg, B. (2016). A Markov decision process model to guide treatment of abdominal aortic aneurysms, *2016 IEEE Conference on Control Applications (CCA), Buenos Aires, Argentina*, pp. 436–441.

Ozcan, Y.A., Tánfani, E. and Testi, A. (2011). A simulation-based modeling framework to deal with clinical pathways, *Proceedings of the 2011 Winter Simulation Conference (WSC), Phoenix, USA*, pp. 1190–1201.

Palumbo, P., Ditlevsen, S., Bertuzzi, A. and Gaetano, A.D. (2013). Mathematical modeling of the glucose–insulin system: A review, *Mathematical Biosciences* **244**(2): 69–81.

Papiez, A., Badie, C. and Polanska, J. (2019). Machine learning techniques combined with dose profiles indicate radiation response biomarkers, *International Journal of Applied Mathematics and Computer Science* **29**(1): 169–178, DOI: 10.2478/amcs-2019-0013.

Schaefer, A., Bailey, M., Shechter, S. and Roberts, M. (2005). Modeling medical treatment using Markov decision processes, *in* M.L. Brandeau *et al.* (Eds), *Operations Research and Health Care*, Springer, Boston, pp. 593–612.

Schwarz, K., Römer, M. and Mellouli, T. (2019). A data-driven hierarchical MILP approach for scheduling clinical pathways: A real-world case study from a German university hospital, *Business Research* **12**: 597–636.

Szwed, P. (2013). Application of fuzzy ontological reasoning in an implementation of medical guidelines, *6th International Conference on Human System Interactions, HSI 2013, Gdańsk, Poland*, pp. 1–10.

Weijters, A., Aalst, W. and Medeiros, A. (2006). *Process Mining with the Heuristics Miner-Algorithm*, Eindhoven University of Technology, Eindhoven.

Xie, X.L. and Beni, G. (1991). A validity measure for fuzzy clustering, *IEEE Transactions on Pattern Analysis and Machine Intelligence* **13**(8): 841–847.

Yadav, P., Steinbach, M., Kumar, V. and Simon, G. (2017). Mining electronic health records: A survey, *arXiv:* 1702.03222.

Yang, X., Han, R., Guo, Y., Bradley, J., Cox, B., Dickinson, R. and Kitney, R. (2012). Modelling and performance analysis of clinical pathways using the stochastic process algebra PEPA, *BMC Bioinformatics* **13** (Suppl 14): S4.

Zhang, Y. and Padman, R. (2016). Data-driven clinical and cost pathways for chronic care delivery, *The American Journal of Managed Care* **22**(12): 816–820.

Zhang, Y., Padman, R. and Patel, N. (2015). Paving the cowpath: Learning and visualizing clinical pathways from electronic health record data, *Journal of Biomedical Informatics* **58**: 186–197.

Rafal Deja is with IBM, Katowice, and WSB University, Dąbrowa Górnicza, Poland. He first graduated in computer science from the Silesian University of Technology, and then completed a postgraduate internship at the University of Milan, Italy, concerned with applying mathematical logic proofs in programming. In 2001, he received his PhD degree in computer science from the Institute of Computer Science, Polish Academy of Sciences, Warsaw. His research interests involve artificial intelligence methods and data mining.

Wojciech Froelich received his Master's degree in computer science from the Gliwice University of Technology, Poland, in 1987. Since 1994, he has been with the Institute of Computer Science, University of Silesia, Sosnowiec, Poland. In 2004, he received his PhD degree in computer science from the AGH University of Science and Technology, Cracow, Poland. In 2017, he received his DSc degree in computer science from the Institute of Computer Science, Polish Academy of Sciences, Warsaw. In 2019, he became an associate professor at the University of Silesia.

Grazyna Deja graduated from the Medical University of Silesia in 1996. She defended her PhD dissertation in 2004 and her habilitation in 2014. She is an associate professor of the Medical University of Silesia, Department of Children Diabetology. As a medical doctor, she has been involved in clinical diabetes care of children for over 20 years. She has been participating in international and national scientific projects concerning numerous aspects of diabetology: genetic, epidemiological, and clinical.

Received: 18 March 2020
Revised: 22 July 2020
Re-revised: 29 September 2020
Accepted: 17 October 2020