Temporal Association Rules for Stratification of Type 2 Diabetic Patients

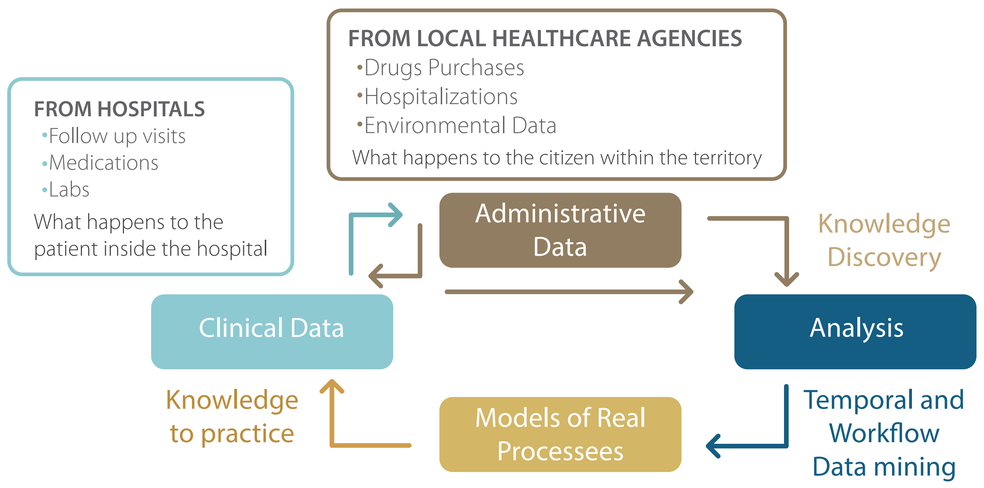
Daniele Segagni[1], Lucia Sacchi[2], Arianna Dagliati[2], Paola Leporati[1], Pasquale Decata[1], John H. Holmes[3], Carlo Cerra[4], Luca Chiovato[1,2], Riccardo Bellazzi[2,1]
1.IRCCS Fondazione S. Maugeri, Pavia, Italy; 2.University of Pavia, Pavia, Italy; 3.University of Pennsylvania, PA; 4.ASL Pavia, Italy

Introduction

In this work we present a knowledge discovery approach to identify significant behavioral patterns that can be used by caregivers to identify the best diagnostic and care pathways for type 2 diabetes (T2DM) patients.

To fill the gaps resulting from infrequent clinical follow-up of diabetic patients, the electronic medical record (EMR) used in IRCCS Fondazione Salvatore Maugeri of Pavia (FSM) has been enhanced with data coming from the local healthcare agency (ASL) of the Pavia area, which contains administrative findings (e.g.,drug prescription/purchase or hospitalization), and with data reporting environmental information (e.g., presence of sports facilities, environmental pollution) provided by the "Regione Lombardia" databases as open data.

By considering the temporal aspects of the evolution of T2DM and its complications, the implemented system offers a novel perspective giving a dynamic picture of the treated population when compared to traditional risk indicators. This work is part of the MOSAIC project, funded by the EU 7th Framework Programme (http://www.mosaicproject.eu/).



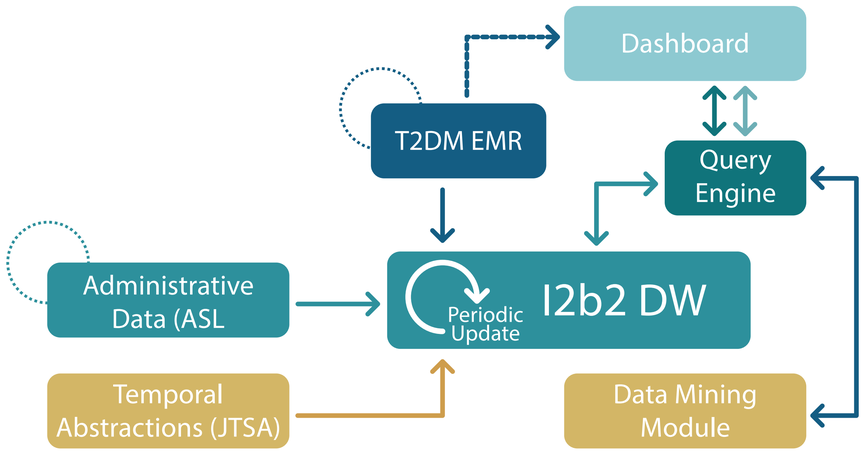
Methods

The core of the system relies on the i2b2 clinical Data Warehouse (DW) installed at FSM. The system contains a module that is able to extract specific abstract concepts from raw quantitative data through the Temporal Abstraction (TA) framework; such concepts are stored in the data warehouse too.

The TA module (JTSA) aims at representing a subset of the clinical and behavioral information in the form of qualitative events. In our implementation, we exploited TAs to represent temporal patterns of glycemic control, weight changes and diet improvements.

The data mining module implements a number of temporal data mining algorithms (namely temporal association rule discovery and sequential pattern mining), which allow extracting and visualizing the most frequent patterns of care the patients are undergoing in the selected population.

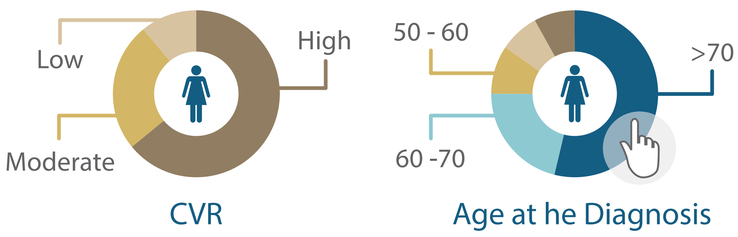
The visualization component of the system is called Dashboard. Through it, doctors can access the DW information and perform analyses during their daily clinical practice. These include visualizing the distribution of the available population based on both raw and abstracted data.



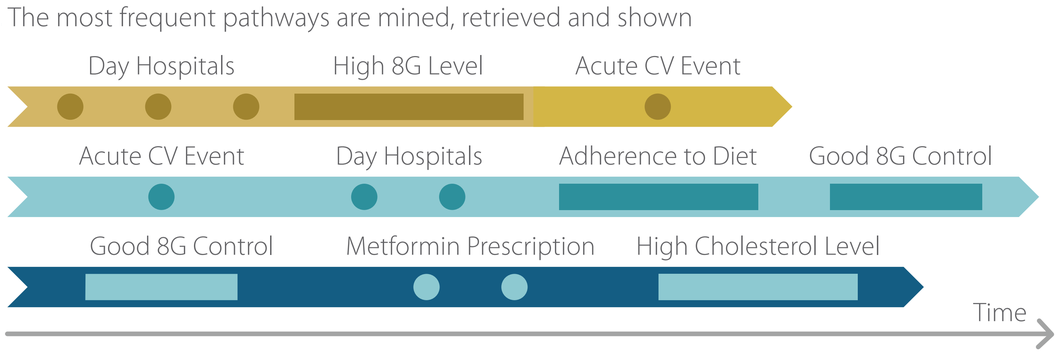
Use Case 1: Hospital Care management

Understanding what is going on in the hospital through a drill down approach: from the whole population to a group of interest.

a. The user can choose a patient set through the distribution of clinical variables in the cohort_

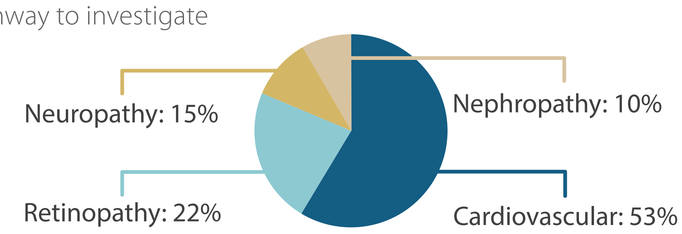


b. The most frequent pathways are mined, retrieved and shown



c. The user can select a specific pathway to investigate

d. The main characteristics of the selected population are displayed

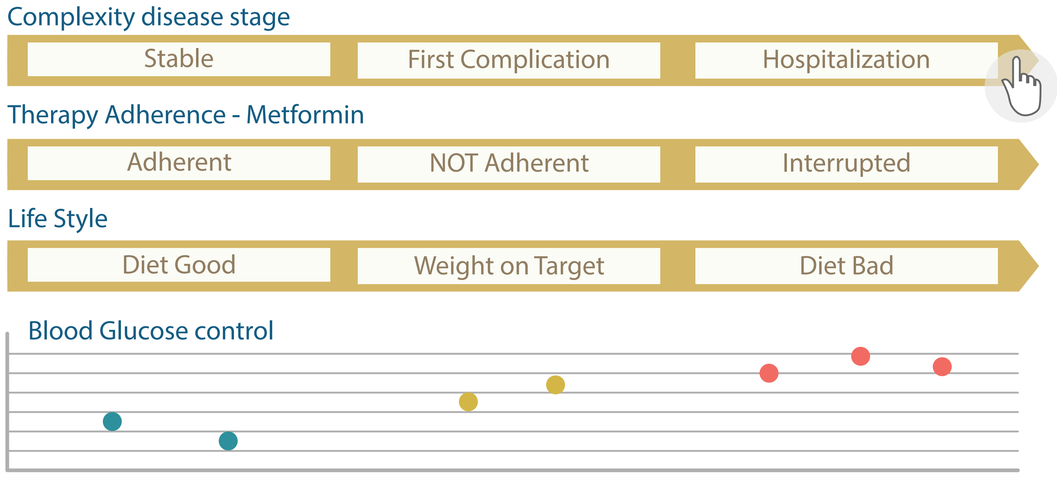


Use case 2: Clinical Decision Support in Follow-up Visits

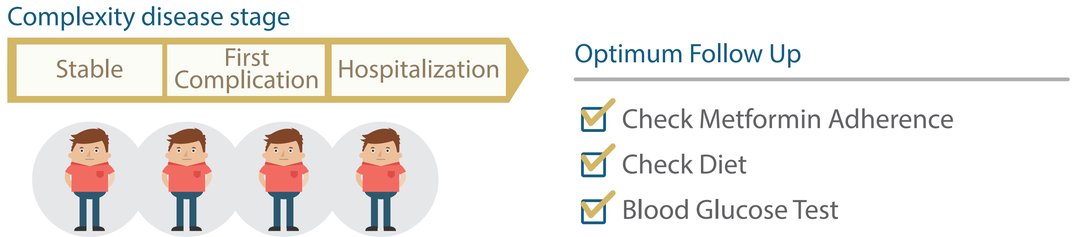
Understanding what is going on at the individual patient level.

a. The user selects a single patient to investigate.

b. Histories of the selected patient are displayed



c. The individual patient history is compared with histories of similar patients to evaluate possible disease evolutions and intervention.



Conclusions

The MOSAIC system offers the opportunity to explore a population of diabetic patients of a specific center and to evaluate its most frequent temporal patterns. These patterns assist in identifying groups of patients with similar disease progressions, giving the possibility of promptly managing the most critical situations and may help decision makers in the allocation of hospital resources in cases of the most demanding patterns.

Co-funded by the Seventh Framework Programme of the European Union

Cardiovascular Risk-Associated Qualitative Pathways in T2D Patients

Arianna Dagliati[1], Lucia Sacchi[1], Daniele Segagni[2], Paola Leporati[2], Luca Chiovato[2,1], Riccardo Bellazzi[1,2].
1.University of Pavia, Pavia, Italy, 2.IRCCS Fondazione S. Maugeri, Pavia, Italy;

Introduction

Clinical data can be integrated with data recorded for administrative purposes to monitor patients' behavior and clinicians' actions. These data streams can be jointly used to show how clinical processes are actually executed. In this context, the main goal of our approach is to stratify the risk of developing T2D-associated complications and to identify significant behavioral patterns that can be used to select the best care pathways for a certain population. To this end, we propose to jointly use temporal data mining and process mining techniques to extract frequent and meaningful healthcare pathways.

We tested this framework on a 1.020 T2D patients data set. This work is part of the MOSAIC EU project, funded by the 7th Framework Program. Administrative data include demographic information, hospital admissions and drug purchases. Clinical variables are related to patho-physiological parameters (BMI, HbA1c, Cholesterol, Blood Pressure) and information about current treatments and life style (Diet, Smoking Habit).

TEMPORAL DATA MINING
- Able to extract frequent patterns from clinical data with temporal features (e.g. TARs)
- Cannot generate complex temporal histories (e.g. chains of events)

PROCESS MINING
- Mines workflows from event logs
- Takes into account only 'process' data (and not quantitative clinical data)

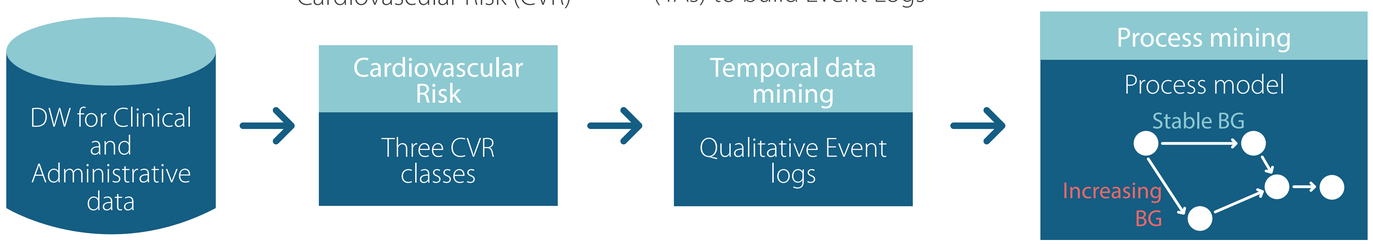
Methods

A. Multivariate Temporal Data stored in a data warehouse (DW)

B. Identification of groups of patients with similar Cardiovascular Risk (CVR)

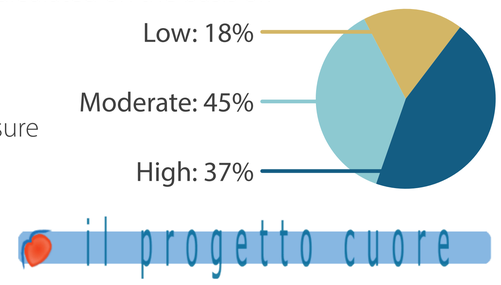
C. Temporal Abstractions (TAs) to build Event Logs

D. Reconstruction of the clinical pathways patients undergo during their process of care (data-aware process mining)



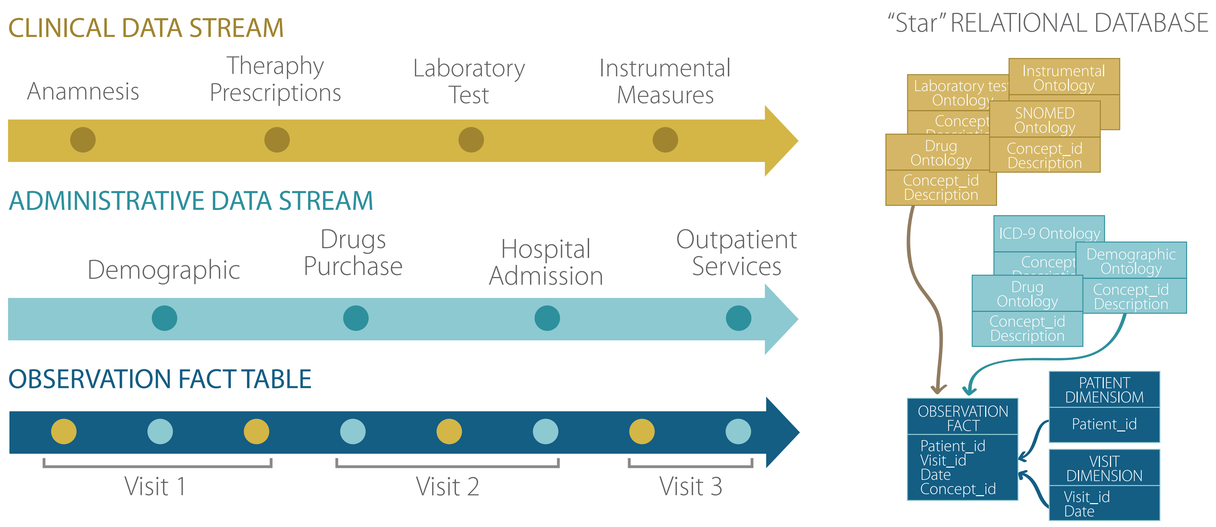
B. Risk Stratification

Three classes of risk calculated on the basis of:
- Age classes
- Gender
- Systolic Blood Pressure
- Total Cholesterol
- Smoking Habit

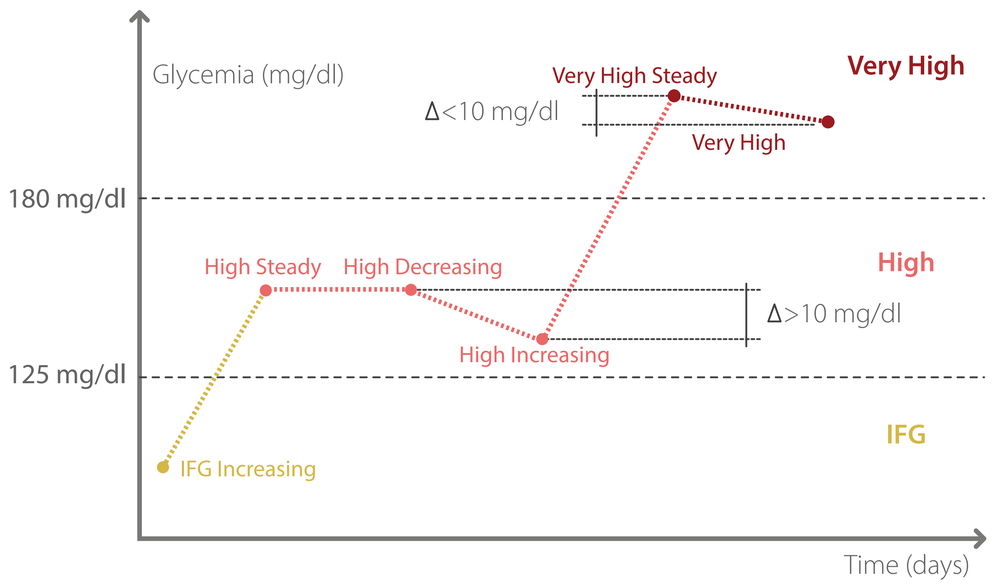


A. Multivariate Temporal Data

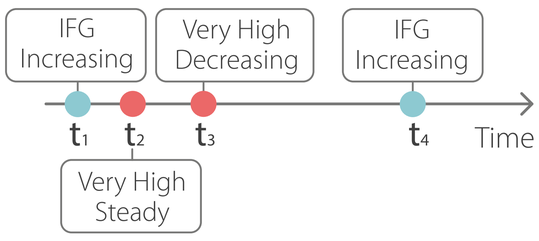
A DW for the collection and integration of multidimensional data from heterogeneous sources



C. TAs to find patterns in Blood Glucose Control



D. Process Mining



- Sequences of Blood Glucose (BG) measures are pre-processed to obtain qualitative TAs.
- Process mining allows extracting the most frequent paths for each CVR class.
- Each step of a BG pattern is enriched with administrative data information on Drug Prescription.

START — **HIGH CVR**

	EVENT	#EVENTS	
A10A	INSULINS AND ANALOGUES	CONTINOUS	10
B01A	ANTITHROMBOTIC AGENTS	CONTINOUS	5
C10A	LIPID MODIFYING AGENTS, PLAN	CONTINOUS	3

HIGH DECREASING

	EVENT	#EVENTS	
A10B	BLOOD GLUCOSE LOWERING DRUGS, EXCL. INSULINS	CONTINOUS	2
A10B	BLOOD GLUCOSE LOWERING DRUGS, EXCL. INSULINS	INTERRUPTION	2
B01A	ANTITHROMBOTIC AGENTS	INTERRUPTION	2

HIGH INCREASING

	EVENT	#EVENTS	
A10B	BLOOD GLUCOSE LOWERING DRUGS, EXCL. INSULINS	INTERRUPTION	2

IFG STEADY

START — **MODERATE CVR**

	EVENT	#EVENTS	
A10B	BLOOD GLUCOSE LOWERING DRUGS, EXCL. INSULINS	CONTINOUS	27
B01A	ANTITHROMBOTIC AGENTS	INTERRUPTION	6
B01A	ANTITHROMBOTIC AGENTS	INTERRUPTION	3
C10A	LIPID MODIFYING AGENTS, PLAIN	CONTINOUS	8

HIGH DECREASING

	EVENT	#EVENTS	
A10B	BLOOD GLUCOSE LOWERING DRUGS, EXCL. INSULINS	CONTINOUS	2
A10B	BLOOD GLUCOSE LOWERING DRUGS, EXCL. INSULINS	INTERRUPTION	2
B01A	ANTITHROMBOTIC AGENTS	INTERRUPTION	2

IFG STEADY

	EVENT	#EVENTS	
A10B	BLOOD GLUCOSE LOWERING DRUGS, EXCL. INSULINS	INTERRUPTION	2

HIGH STEADY

START — **LOW CVR**

	EVENT	#EVENTS	
A10B	BLOOD GLUCOSE LOWERING DRUGS, EXCL. INSULINS	CONTINOUS	9
C10A	LIPID MODIFYING AGENTS, PLAIN	CONTINOUS	6

VERY HIGH DECREASING

	EVENT	#EVENTS	
A10A	INSULIN AND ANALOGUES	INTERRUPTION	1
A10B	BLOOD GLUCOSE LOWERING DRUGS, EXCL. INSULINS	CONTINOUS	2

VERY HIGH INCREASING

	EVENT	#EVENTS	
A10B	BLOOD GLUCOSE LOWERING DRUGS, EXCL. INSULINS	CONTINOUS	6
C10A	LIPID MODIFYING AGENTS, PLAIN	CONTINOUS	2

HIGH INCREASING

Results and Conclusions

Result: One of the most interesting results comes from the comparison of High and Low risk classes. When histories start with very high Blood Glucose values, in the Low risk class they mainly end in the same state or worse, while in the High risk class is possible to recognize improvement paths in more than half cases. **Conclusion:** This work tackles the major challenges we faced managing complex clinical and administrative temporal data so to identify and compare clusters of relevant healthcare pathways.

Co-funded by the Seventh Framework Programme of the European Union

Exploring Continuous Glucose Monitoring on the frequency domain to identify risk factors in type 2 Diabetes

Miguel Maria Isabel(1), Jorge Cancela(1), Giuseppe Fico(1), Andrea Facchinetti(2), Chiara Fabris(2), Claudio Cobelli(2), Maria Teresa Arredondo(1) , on behalf of the MOSAIC Consortium.
(1)Life Supporting Technologies, Universidad Politecncia de Madrid - Madrid, Spain. (2)Department of Information Engineering, University of Padova – Padova, Italy

Introduction

Glucose variability (GV) is believed to be a key indicator of risk factors in both individuals with Type 1 (T1DM) and Type 2 Diabetes Mellitus (T2DM). Continuous Glucose Monitoring (CGM) devices allow a better characterization of GV thanks to the almost continuous monitoring of glucose concentration. A review of the literature revealed that all GV indices derived from CGM data are calculated on the time domain, while the frequency domain is still unexplored. This work is aimed to assess the CGM signal on frequency domain trying to identify new features for a better characterization of T2DM.

This work is part of the MOSAIC project, which is an EU-funded European ICT project carried out within the 7th Framework Program devoted to the development of mathematical models and algorithms that can enhance the current tools and standards for the diagnosis of T2DM, IGT and IFG; that can improve the characterization of patients suffering those metabolic disorders and that can help evaluating the risk of de veloping T2DM and their related complications.

Motivation: CGMS State-of-the-art

38 papers meet the inclusion criteria (17 were excluded). Research topics identified:

- Comparison between different CGMS devices (accuracy, reliability or lifetime)
- Methods for interpreting data
- Closed-loop system. How to introduce a CGMS in a closed-loop.
- Study the relationship between diabetes and other diseases
- Tracking the evolution of different treatments based on CGMS signal.
- Early diagnosis of diabetes disease. People in risk i.e. obese people with first-degree relatives of T2DM.
- Looking for new parameters to study GV (glycaemic variability).

- The research evidenced that GV is proposed as the most important measurement in the assessment of diabetes.
- The analysis highlighted that SDT is the most used glycaemic variability index, which was found to be used in approximately half of studies. The other largely used GV indices are MBG and MAGE.
- However, **it is important to evidence that a gold standard index to access GV is still not available.**
- Several indices are strongly correlated each other, suggesting that the use of many of them is redundant.
- Finally, it can be observed that GV is mainly studied in the time domain. **A research proposal is to extend the study of the signal and its variability to the frequency and to the time-frequency domains.** Exploring these representations could lead to the identification of new risk factors for T1DM and T2DM.

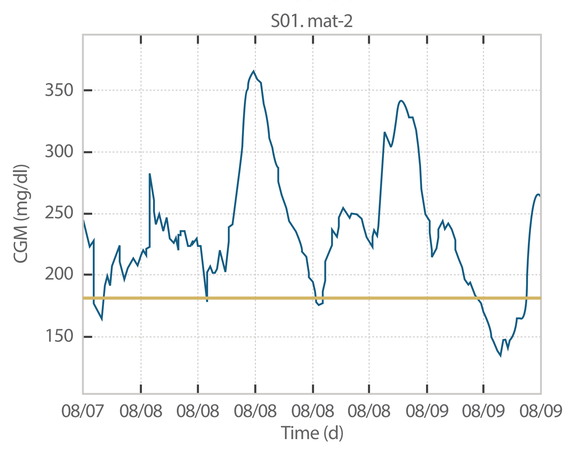
Methods

The proposed frequency analysis consists in transforming CGM signals to the frequency domain using the Fast Fourier Transform (FFT) and mining several spectrum parameters: the maximum power spectral density (PSD) and the frequency where it is located, PSD of the second spectrum peak and the its relative frequency, difference between the maximum and the second peak of the spectrum.

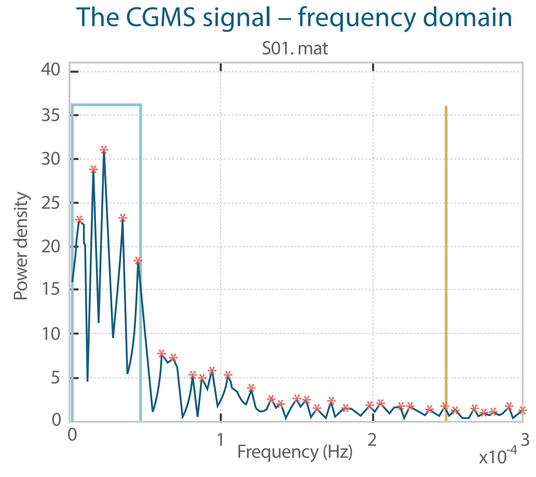
Proposal of frequency indices

- **Maximum PSD and the frequency where is located**
  This parameter shows the value of the maximum peak of the spectrum and the frequency where is located.
- **PSD of the second spectrum peak and the frequency where is located**
  In that case, the value of the second peak of the spectrum and their location are shown.
- **Difference between the maximum and the second peak of the spectrum**
  This parameter shows the PSD difference between the two parameters above. It is the difference between maximum PSD of the spectrum and the PSD of the next peak below.
- **3dB bandwidth related to the maximum spectrum peak and the cut frequencies**
  First, the value of the 3dB bandwidth related to the maximum spectrum peak is shown, and then the frequencies among which is.
- **Frequency where PSD reaches the percentage defined**
  Finally, this parameter is related to the last graphic parameter explained. It shows the value of the frequency where the spectrum reach a defined percentage of the total spectrum power.

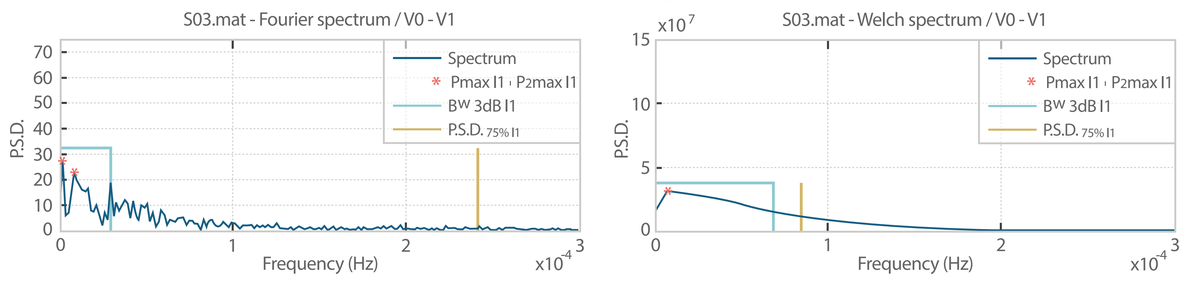
Method



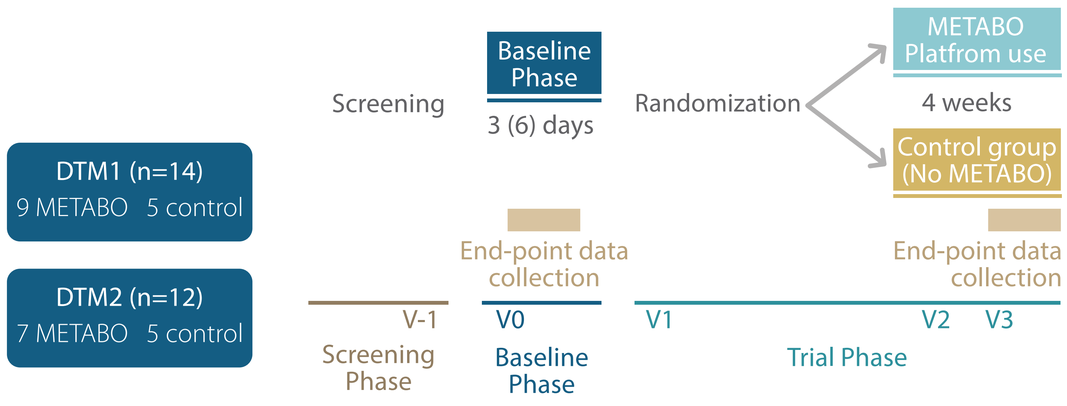
The CGMS signal – frequency domain



FFT vs. Welch



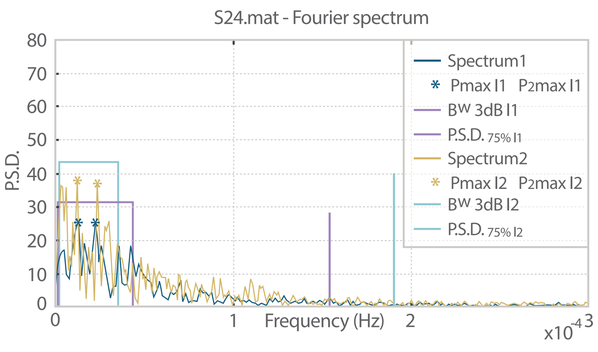
Data collection



Visit 1 vs Visit 2

This figure is an example of the frequency signal before and after the intervention. It also shows visually some of the frequency features calculated from the frequency domain such as: frequency and value of the maximum peak and the second peak, band which (at -3dB), and the frequency encompassing 75% of the Power Spectral Density.

Data were collected before and after the intervention in the METABO group.



Results and Conclusions

Result: The method was applied on 11 T2DM subjects recruited during the METABO research project monitored with CGM for 3 days in two sessions 1-month apart, before and after the intervention. Comparison of power spectra evidenced narrower and higher spectra after the intervention. This is in line with expectations, being the spectrum of a regular signal (i.e. with reduced GV) less wide.
Conclusion: This introductory study evidenced that frequency analysis of CGM data is a candidate tool for the evaluation GV. Further work should be carried out on the validation of frequency parameters as GV indexes, e.g. exploiting the upcoming CGM datasets that will be collected within the MOSAIC project funded by the EU under the FP7 framework.

Co-funded by the Seventh Framework Programme of the European Union