Remote Accessibility to Diabetes Management and Therapy in Operational healthcare Networks.

REACTION (FP7 248590)

D6.2
Newly discovered diabetes knowledge in publicly available and REACTION datasets

Date 2012-02-29

Version 1.1

Dissemination Level: Public

Table of Content

Document control page

Code	D6.2_Newly discovered diabetes knowledge in publicly available and REACTION datasets_V1.1.doc
Version	1.1
Date	2012-02-29
Dissemination level	P
Category	R
Participant Partner(s)	MSG, FORTH, BTS, ALL
Author(s)	Stephan Spat, Vincenzo Lagani, Stephan Schaller, Tamas Gergely, Ioannis Tsamardinos, Tamás Tóth, Miklós Szőts, András Simonyi
Verified and approved by	
Work Package	WP6
Fragment	No
Distribution List	All
Abstract	This deliverable gives an insight into the model development for long and short-term risk predication for primarly patients with T1DM, and the work on insulin dosing for non-ICU T2DM patients. Publicly available diabetes datasets and path-breaking literature for non-ICU T2DM treatment have been used as basis for this work.
Comments and modifications	
Status	☐ Draft ☒ Task leader accepted ☒ WP leader accepted ☐ Technical supervisor accepted ☐ Medical Engineering supervisor accepted ☐ Medical supervisor accepted ☐ Quality manager checked ☐ Project Coordinator accepted
Action requested	☐ to be revised by partners involved in the preparation of the deliverable ☐ for approval of the task leader ☐ for approval of the WP leader ☐ for approval of the Technical Manager ☐ for approval of the Medical Engineering Manager ☐ for approval of the Medical Manager ☐ for approval of the Quality Manager ☐ for approval of the Project Coordinator **Deadline for action**: N/A
Keywords	Risk assessment, risk models, diabetes data sets, insulin dosing protocol, text mining
References	
Previous Versions	

Version Notes

Version	Author(s)	Date	Changes made
0.1	S. Spat	2011-10-19	TOC
0.2	S. Spat	2011-12-06	Update content
0.3	S. Spat	2012.01-19	TOC refactoring
0.4	T. Toth, V. Lagani, S. Spat, S. Schaller, M. Szőts, A. Simonyi, Tamas Gergely, Ioannis	2012-02-22	Content provision for chapters, update of document structure

		Tsamardinos,		
	1.0	S. Spat, T. Toth, S. Schaller, V. Lagani	2012-02-23	Finalization
	1.1	S. Spat, V. Lagani	2012-02-29	Update according to reviews comments
Internal review history	Reviewed by		Date	Comments made
	Giorgos Vasilakis		2012-02-27	Approved with comments
	Peter Rosengren		2012-02-25	Approved with comments

1. Executive Summary

D6.2 focuses on discovering new knowledge for the treatment of diabetes from publicly available and REACTION datasets in the context of risk predication and insulin dosing. The following work has been performed for four different fields of applications within WP6:

Long-term risk modelling: Diabetes increases the probability of developing several serious health problems; a correct evaluation of the long term risk can help both physicians and patients in devising more effective personalized health care plans. FORTH developed long term risk assessment models for six different diabetes related complications: Severe Hypoglycaemia, ketoacidosis, neurobehavioral complications, nephropathy, neuropathy and retinopathy. The models were derived by employed a state-of-the-art, multivariate data analysis approach on the data collected in the Diabetes Control and Complications Trial (DCCT), a ten-years long, randomized case-control study for T1DM. Five out of six models showed a high predictive performance, with a nested-cross validated Concordance Index ranging from 0.67 to 0.79. Moreover, the models provided interesting insights about the risk factors more related to each complication.

Physiology-based glucose-insulin model: The model kernel developed by BTS is, among other applications, used to support short-term risk assessment and model-based data analysis and inference. However, the strategy chosen for the development of the model kernel takes into consideration that the area of application of the model is multi-facetted. Besides risk assessment and model-based data analysis for knowledge generation, which necessitate the possibility to integrate detailed and specific knowledge on physiological conditions of the patients with diabetes, the model will also be used for automated blood glucose control in patients with (T1DM and T2DM) diabetes. Consequently, a detailed mechanistic modelling approach using physiology-based pharmacokinetic / pharmacodynamics (PBPK/PD) model kernels was chosen. The models are characterized by a detailed description of the human organism and allow vertical integration of information on multiple scales. Using the modelling tools by Bayer Technology Services, these models allow data integration ranging from a molecular scale, e.g. cellular signalling cascades to population scale, e.g. the distribution of anthropometric data within a patient cohort. Importantly, the approach and the tools allow model individualization. Thus, the physiology-based model kernel provides a powerful basis for short-term risk analysis and the generation of new knowledge using heterogeneous input data from different physiological scales. Model-based analysis of data from clinical trials conducted at the REACTION partner MUG indicates that insulin kinetics shows over-proportional clearance rates in the lower concentration range, which could be explained by allosteric binding of the insulin receptor, and further, that dynamic change in insulin clearance observed after large insulin boluses could be caused by feedback-loops within the receptor model.

Insulin dosing support for hospital: Hyperglycaemia is associated with poor clinical outcome and increased mortality. In the hospital the aim is to cure the primary condition that caused the hospitalization and to keep other complications within certain limits which results very often in poor glycaemic control. In this deliverable MSG and MUG present the results of the development of insulin dosing protocol used at general wards for patients with diabetes type 2. The dosing protocol is embedded into a mobile tablet-based workflow and decision support system. A team of physicians, nurses and engineers designed the user interface, the functionalities and the REACTION protocol. The REACTION-protocol is based on the basal/bolus-protocol of the RABBIT-2 trials of Umpierrez et al. (2009) which proved to be most promising for our purpose. In the final step, we implemented the user requirements and findings of the protocol reviews in a software prototype according to the standards based on the Medical Device Directive. Currently the REACTION protocol is validated on paper in a clinical study. In the next step the electronic glucose management system including the protocol will be tested in a clinical study at the Medical University of Graz.

Text mining for decision support in diabetes treatment: ALL has developed a semantic information retrieval technology which does not require a comprehensive domain-specific ontology in order to achieve high precision values. Our aim is to implement this technology to be used in the diabetes domain within the REACTION platform. The developed component will support its users in finding the most appropriate information for example in electronic patient records, guidelines or evidence-based literature repositories. The main elements of the solution are discussed and a detailed work plan is given in this deliverable. The work on generic elements has already been started. Input from potential users is necessary in order to gather their requirements and tailor the solution to the actual needs. The prototype of this component is expected by the end of year 3.

2. Abbreviations

AFT	Accelerated Failure Models
ALL	Applied
ANS	Autonomous Nervous System
BTS	Bayer Technologies
BVS	Bayesian Variable Selection
CHC	Chorleywood Health Centre
CHD	Coronary Heart Disease
CI	Concordance Index
CSII	Continuous Subcutaneous Insulin Infusion
CVD	Cardiovascular Disease
DAG	Directed Acyclic Graph
DCCT	Diabetes Control and Complications Trial
EDIC	Epidemiology of Diabetes Interventions and Complications
EPR	Electronic Patient Record
ESRD	End Stage Renal Disease
FORTH	Foundation for Research and Technology - Hellas
GUI	Graphical User Interface
HbA1c	Glycosylated Hemoglobin, Type A1C
HDL	High Density Lipoprotein
IE	Information Extraction
IR	Information Retrieval
IRS-1	Insulin Receptor Substrate-1
LDL	Low-Density Lipoprotein (Cholesterol)
MeSH	Medical Subject Headings
MI	Myocardial Infarction
MPC	Model Predictive Control
MSG	Institute for medical technologies and health management (now HEALTH)
NER	Named Entity Recognition
NICE	National Institute for Clinical Excellence
NIDDK	National Institute of Diabetes and Digestive and Kidney Disease
NIDDM	Non-Insulin Dependent Diabetes Mellitus
NLP	Natural Language Processing
ICU	Intensive Care Unit
PBPK	Physiology-based Pharmacokinetic
PD	Pharmacodynamics
PKC	Protein Kinase C
RSF	Random Survival Models

SIR	Semantic Information Retrieval
SMMPC	Survival Max Min Parents and Children
SSI	Sliding Scale Regular Insulin
SVCR	Support Vector Approach to Censored Targets
SVM	Support Vector Machines
T1DM	Type 1 Diabetes Mellitus
T2DM	Type 2 Diabetes Mellitus
UDPK	Undiscovered Public Knowledge
UKPDS	United Kingdom Prospective Diabetes Study

3. Introduction

3.1 Purpose, context and scope of this deliverable

Based on the work of D6.1 “Disease Management Strategies for Diabetes”, which described existing disease management strategies and available predicative risk models and multi-parametric risk assessment methods from literature, D6.2 highlights the realization and results of risk modelling and insulin dosing protocol development in the context of the work on REACTIONs risk engine.

The purpose of this deliverable is to present the concepts and first results of risk modelling to support the management of diabetes in primary care as well as at the general wards. Depending on the end user needs and on publicly available data sets, long-term risk and physiology models for T1DM and an insulin dosing protocol for T2DM have been developed. In particular, D6.2 focuses on newly discovered knowledge about diabetes from publicly available datasets which will be used for risk predication in short- and long-term complications. In addition, a concept using text mining methods for medical decision support is presented.

3.2 Structure of the document

Four different fields of application can be distinguished according to the underlying problems and the proposed modelling:

- Long-term risk predication to support medical decision making and patient communication in primary care
- Physiology-based modelling for hypoglycaemia detection
- Insulin dosing protocol to improve diabetes treatment of non-ICU T2DM patients in hospital
- Knowledge-generation from medical literature to link patient profile to treatment relevant literature using text mining

Chapter 4 starts with the description of the long-term risk modelling. Six different risk predication models for T1DM will be presented including the design and discovered knowledge in the publicly available DCCT dataset. Chapter 5 presents the work on physiology-based glucose-insulin modelling. An overview of state of the art is followed by a description of the methods and main outcomes based on a diabetes dataset from routine care provided by the Medical University of Graz (MUG). Chapter 6 focuses on the hospital domain. An insulin dosing protocol for the improvement of diabetes management at a general ward has been developed based on published literature and has been used as the basis for a mobile tablet-based workflow and decision support system. Finally, chapter 7 gives an outlook how text mining technologies can support for knowledge discovering in diabetes. A concept for using a semantic information retrieval technology in the diabetes domain will be presented.

Before discussing the development work, a short overview regarding diabetes-related complications and risk factors is given in the next sections.

3.3 Complications and their risk factors in diabetes

Diabetes has both short-term and long-term effects and complications. The former are acute, often life-threatening problems which require immediate treatment. Long-term complications develop gradually over the years, and without intervention may lead to severe problems. If short-term problems are not properly treated, they accelerate the development of long-term complications. T2DM mostly begins without symptoms. Therefore, the complications may be present already at the time of diagnosis.

3.3.1 Short-term complications

The short-term complications are usually consequences of inadequate treatment. The most common conditions include:

- **Hypoglycaemia**: can occur in insulin-treated patients if too much insulin is injected. Physical activity, missed meals and certain medications may also cause hypoglycaemia. If the patient remains conscious, sugar-rich food or drink should be consumed. Otherwise, a glucagon injection is required.

- **Ketoacidosis**: it is more common in T1DM patients and often the first sign of diabetes. Excessive insulin deficiency leads to lipolysis (i.e. breakdown of fats) in the adipose tissues which increases the number of ketone bodies in the blood. This causes frequent urination and dehydration. Ketones can be detected in the urine. This condition requires hospital treatment. Without treatment it may cause coma.
- **Hyperosmolar Non Ketotic Coma**: it occurs mainly by T2DM patient. An extremely high blood sugar level leads to excessive urination and dehydration and finally coma. It requires immediate treatment in hospital.
- **Mental problems**: A study conducted by University of Virginia showed that hyperglycaemia in diabetic patients can slow down brain function and reduce cognition. Verbal ability is impaired, as is the ability to perform simple tasks. Resolution is achieved by restoring a normal blood sugar.
- **Infections**: Diabetic patients have increased susceptibility to various infections such as: tuberculosis, pneumonia, pyelonephritis, carbuncles and diabetic ulcers. This may be due to: poor blood supply reduced cellular immunity, and hyperglycaemia. Some of them can be prevented by vaccination.
- **Skin problems**: Diabetic dermopathy appears as roundish and slightly indented patches of skin that are brown or purplish in colour. It appears to be linked to earlier trauma and slow healing. It is associated with high glucose levels and normalization of glycaemic control helps to resolve this problem. Fungal infections on several parts of the body are commoner in the diabetic patient, especially those that are poorly controlled.

3.3.2 Long-term complications

Uncontrolled (high) blood sugar damages the blood vessels and causes most of the long-term complications. Damage to large vessels causes macrovascular complication and damage of small vessels (capillaries) causes microvascular complications.

Macrovascular complications

- **Cardiovascular diseases**: Diabetes may cause fatal complications such as coronary heart disease (leading to a heart attack). People with diabetes are two to four times more likely to develop cardiovascular disease than people without diabetes. Diabetic patients may suffer silent Heart Attacks that can occur without classical symptoms. It is believed that a diabetic patient is unaware because nerve damage prevents the transmission of pain.
- **Cerebrovascular diseases**: Damage to cerebral arteries causes stroke that can result in death or impaired brain function. The risk of stroke in the diabetic patient is more than double the rate for the general population. A study conducted by Harvard University showed that diabetes can impair memory and cause a permanent loss of cognitive function over time. It can also lead to Alzheimer's disease.

Microvascular complications

The damage of the capillaries may affect several organs causing severe complications. The most common conditions include:

- **Neuropathy**: damage to the nerve fibres primarily affecting the legs and feet. Foot ulcers occur commonly. Infections in these wounds may ultimately result in amputation of the foot and lower leg. It is estimated that up to 70% of all lower limb amputations are related to diabetes. There are various types of diabetic neuropathy: peripheral, autonomic, proximal, and focal. Common symptoms of diabetic neuropathy are numbness, tingling, decreased sensation to a body part, diarrhoea, constipation, loss of bladder control, impotence, facial drooping, drooping eyelid, drooping mouth, vision changes, weakness, speech impairment, etc. These symptoms usually develop gradually over years.
- **Nephropathy**: may result in total kidney failure and in the need for dialysis or kidney transplant. Diabetes is the leading cause of kidney failure in the developed world and accounts for approximately 35 to 40 % of new cases of End Stage Renal Disease (ESRD) each year. Microalbuminurea is an early sign of kidney damage.

- **Retinopathy**: damage of the retina of the eye can lead to vision loss. The incidence of blindness is 25 times higher in people with diabetes than in the general population. Cataracts are also common is diabetic patients.

3.3.3 Factors for the risk of complications

Common characteristics of risk factors:

- The condition must be associated with the disease in a manner which is beyond chance alone. A causal link is therefore implied.
- A risk factor will not necessarily always lead to the development of the disease.
- The ultimate purpose of identifying a risk factor is to modify it in order to prevent the disease.

Types of risk factors

- **Non-modifiable risk factors**: factors like age and sex influence the risk of some conditions but they cannot be modified.
- **Genetic:** it is often costly, inconvenient, or impossible to measure directly to date, but family history of certain conditions (e.g. coronary heart disease before 55 years of age) may indicate increased risk. Diabetes is also more common in certain ethnic groups; that implies genetic background. T1DM has a significant genetic component. Inheritance complex and not fully understood and has yet to be clinically valuable. T2DM appears to have a number of forms. Some of these also have a genetic determination indicated by familial characteristics and a significant relationship between birth weight and later development of T2DM and its complications.
- **Lifestyle**: it has a great impact on the development of diabetes and its associated conditions. The main lifestyle risk factors are the following:
    - *sedentary lifestyle*
    - *unhealthy diet*
    - *smoking*
    - *stress*
    - *alcohol*: moderate alcohol consumption may lower risk, but excessive consumption is a risk factor
- **Physiological parameters**: some risk factors can be identified as single measurements.
    - *High blood glucose levels:* poor control of diabetes is one the most important risk factors. There is conclusive evidence that good control of blood glucose levels can substantially reduce the risk of developing complications and slows the progression of all types of diabetes.
    - *High blood pressure:* it is an important risk factor for both microvascular and macrovascular complications.
    - *Dyslipidaemia (High cholesterol, high triglycerides, low LDL):* it is another important risk factor of cardiovascular diseases.
- **Co-morbidities**
    - *Obesity*: it is mostly caused by an unhealthy lifestyle. It is a risk factor for both T2DM and several long-term complications. It is especially central obesity that is associated with diabetes and its complications.
    - *Hypertension*: Prevalence is at least double in people with T2DM.
    - *Left ventricular hypertrophy:* Most commonly seen in people with long-standing high blood pressure, but is also seen in the absence of elevated blood pressure in people with diabetes.

4. Long-term risk modelling

4.1 State-of-the-art and related work

During the last decades medical researchers have given increasing attention to the study of diabetes complications risk factors. A practical application of risk factors studies is found in the development of risk assessment models. These models are able to provide an estimation of the patient's risk of developing diabetes complications, given the patient's profile. In particular, several models have been introduced for assessing the *long term* risk of developing complications, or experiencing adverse events related to diabetes.

A detailed description of available long term risk assessment models and related studies is provided in the deliverable D6.1 "Disease Management Strategies for Diabetes". Here we report short versions of the D6.1 review.

4.1.1 Qrisk study

A large study in the field of diabetes risk factor is the QRisk [Hippisley-Cox et al. (2007)]. QRisk aims to develop a cardiovascular disease risk algorithm which will provide accurate estimates of cardiovascular risk in patients from different ethnic groups in England and Wales and to compare its performance with the modified version of the Framingham score recommended by the National Institute for Health and Clinical Excellence (NICE). 2.3 million Patients aged 35-74 with 140 000 cardiovascular events participated in the QRisk study. Overall population (derivation and validation cohorts) comprised 2.22 million people who were white or whose ethnic group was not recorded, 22 013 south Asian, 11 595 black African, 10 402 black Caribbean, and 19 792 from Chinese or other Asian or other ethnic groups.

Current version of the calculator is QRisk2 and can be found online at (http://www.qrisk.org/). QRisk2 uses the following parameters: age, gender, current smoker (yes/no), family history of heart disease aged <60 (yes/no), existing treatment with blood pressure agent (yes/no), postcode (postcode-related Townsend score) - an area measure of deprivation, body mass index (height and weight), systolic blood pressure (use current not pre-treatment value), total and HDL cholesterol, self-assigned ethnicity, rheumatoid arthritis, chronic kidney disease and atrial fibrillation. QRisk2 uses Cox proportional hazards models in the derivation dataset to estimate the coefficients and hazard ratios associated with each potential risk factor for the first ever recorded diagnosis of cardiovascular disease for men and women separately. QRisk2 uses fractional polynomials to model non-linear risk relations with continuous variables, where appropriate, and tests for interactions between each variable and age and between diabetes and deprivation.

4.1.2 UKPDS study

Another online risk calculator for people with T2DM is the UKPDS risk engine (http://www.dtu.ox.ac.uk/riskengine/). The U.K. Prospective Diabetes Study [UKPDS Group (1991)] (UKPDS) is a landmark randomized controlled trial which showed that both intensive treatment of blood glucose and of blood pressure in diabetes can lower the risk of diabetes-related complications in individuals newly diagnosed with T2DM. The UKPDS cohort consists of 5102 patients, followed for a median of 10.7 years. Between 1977 and 1991, general practitioners in the catchment areas of 23 participating UKPDS hospitals were asked to refer all patients aged 25 to 65 years presenting with newly diagnosed diabetes. Patients in the UKPDS had biochemical measurements, including HbA1c, blood pressures, and lipid and lipoprotein fractions, recorded at entry to the study, at randomization in the study after a three-month period of dietary therapy, and each year subsequently.

The UKPDS Risk Engine [Stevens et al. (2001)] provides risk estimates and 95% confidence intervals, in individuals with T2DM not known to have heart disease, for (a) non-fatal and fatal coronary heart disease, (b) fatal coronary heart disease, (c) non-fatal and fatal stroke and (d) fatal stroke. These can be calculated for any given duration of T2DM based on current age, sex, ethnicity, smoking status, presence or absence of atrial fibrillation and levels of HbA1c, systolic blood pressure, total cholesterol and HDL cholesterol.

[Stevens et al. (2004)] using a subset of the UKPDS cohort enabled estimation of the probability of fatal coronary heart disease (CHD) and fatal stroke within the UKPDS Risk Engine or other computer models. The analysis was based on 674 cases of myocardial infarction MI (351 fatal) that occurred in

597 out of 5,102 UKPDS patients for whom covariate data were available during a median follow-up of 7 years.

Another publication based on the UKPDS cohort is the [Kothari et al. (2002)]. This study proposed mathematical models to estimate the risk of a first stroke using data from 4549 newly diagnosed T2DM patients enrolled in the UKPDS Study. This model forecasts the absolute risk of a first stroke in people with T2DM using variables readily available in routine clinical practice.

[Clarke et al. (2004)] developed the UKPDS Outcomes Model for T2DM that can be used to estimate the likely occurrence of major diabetes related complications over a lifetime. Equations for forecasting the occurrence of seven diabetes-related complications and death were estimated using data on 3642 patients from the United Kingdom Prospective Diabetes Study. The model's forecasts fell within the 95% confidence interval for the occurrence of observed events during the UKPDS follow-up period. When the model was used to simulate event history over patients' lifetimes, those treated with a regimen of conventional glucose control could expect 16.35 undiscounted quality-adjusted life years, and those receiving treatment with intensive glucose control could expect 16.62 quality-adjusted life years.

4.1.3 EuroDiab study

Another European diabetes study is the EURODIAB IDDM Complications Study. EuroDiab is a cross sectional study which examined 3,250 T1DM patients. Participants were aged between 15 and 60 years and recruited from 31 centres in 16 European countries. The sampling frame was all T1DM attending at least once in the last year for each centre. Patients were stratified by age (three categories), diabetes duration (three categories), and sex. Patient measurements were taken at baseline (1990–1991) and at 7 years follow-up (1997–1999).

[Vergouwe et al. (2010)] used a subset of the European Diabetes Prospective Complications Study (n = 1115) to develop and validate a clinical prediction rule that estimates the absolute risk of microalbuminuria. Logistic regression was used to estimate multivariable regression coefficients and odds ratios with 95% CIs for each predictor. The number of predictors was reduced with backward stepwise selection. Finally the logistic model was transformed in a risk chart for making the risk calculation easier in real medical settings.

[Skevofilakas et al. (2010)] created a decision support system able to predict the risk of a T1DM patient to develop retinopathy using the EuroDiab baseline dataset. The decision support system is a hybrid infrastructure combining a Feed forward Neural Network, a Classification and Regression Tree and a Rule Induction C5.0 classifier, with an improved Hybrid Wavelet Neural Network. A voting mechanism is utilized to merge the results from the four classification models.

4.1.4 Cleveland study

Another study related to diabetes and its complications is based on the Cleveland Clinic. [Wells et al. (2008)] created a tool that predicts the risk of mortality in patients with type 2 diabetes. This study was based on a cohort of 33,067 patients with T2DM identified in the Cleveland Clinic electronic health record and were initially prescribed a single oral hypoglycemic agent between 1998 and 2006. Follow-up in the cohort ranged from 1 day to 8.2 years (median 28.6 months), and 3,661 deaths were observed. Mortality was determined in the EHR and the Social Security Death Index. A prediction tool was created using the Cox model coefficients. The tool was internally validated using repeated, random subsets of the cohort, which were not used to create the prediction model. The coefficients from the fitted Cox model were also used to develop an interactive web based calculator which available from http://www.clinicriskcalculators.org.

4.1.5 Sweden study

Another independent study can be found at the Swedish National Diabetes Register [Cederholm et al. (2008)].The study is based on 11,646 female and male patients, aged 18–70 years, from the Swedish National Diabetes Register with 1,482 first incident CVD events on 58,342 person-years with mean follow-up of 5.64 years. This study presents a diabetes-specific equation for estimation of the absolute 5-year risk of first incident fatal/nonfatal cardiovascular disease (CVD) in type 2 diabetic patients. All predictors included were associated with the outcome ($P < 0.0001$, except for BMI $P = 0.0016$) with Cox regression analysis. Calibration was excellent when assessed by comparing observed and predicted risk. Discrimination was sufficient, with a receiver operator curve statistic of 0.70. Mean 5-year risk of CVD in all patients was 12.0±7.5%, whereas 54% of the patients had a 5-year risk >=10%.

4.2 Multi-parametric risk assessment for the REACTION risk engine

For the creation of the long term risk assessment models of the REACTION project we required and obtained the DCCT/EDIC dataset from the National Institute of Diabetes and Digestive and Kidney Disease. The characteristics of the DCCT/EDIC study are summarized in Section 4.2.1. The derivation of the risk assessment models furthermore required the adoption of a principled experimentation protocol, for selecting the subset of variables with the highest predictive value while avoiding overfitting of the data. We thus closely followed the protocol reported in [Lagani and Tsamardinos (2010)], that employs a nested cross validation procedure coupled with a permutation based test for identifying the best set of predictors, along with the maximally predictive model. Details of the experimentation are reported in the next sections.

4.2.1 DCCT/EDIC study

The Diabetes Control and Complications Trial (DCCT) is one of the most well-known study of long term risk assessment related to diabetes complications. The DCCT consists in a landmark medical study conducted by the United States National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK). The DCCT involved 1,441 volunteers, ages 13 to 39, with T1DM and 29 medical centres in the United States and Canada. DCCT is a multicentre, randomized clinical trial designed to compare intensive with conventional diabetes therapy with regard to their effects on the development and progression of the early vascular and neurologic complications. Volunteers had to have had diabetes for at least 1 year but no longer than 15 years. The study compared the effects of standard control of blood glucose versus intensive control on the complications of diabetes. Intensive control meant keeping haemoglobin A1C levels as close as possible to the normal value of 6% or less. A new study started after the DCCT, called Epidemiology of Diabetes Interventions and Complications (EDIC) [EDIC Group (1999)]. EDIC is a follow up study on 90% of the participants from DCCT that looked into cardiovascular disease and the effects of intensive control on quality of life and cost effectiveness.

4.2.2 Initial selection of predictors and outcomes definition

DCCT data needed to undergo a phase of data preparation and pre-processing, before proceeding with the derivation of the long term risk assessment models. In particular, it is necessary to define (a) the clinical parameters to be employed as predictors through our analysis, and (b) the outcome to be considered:

a) We selected 80 different parameters measured at baseline as initial set of predictors for the derivation of the long term risk assessment models. The complete list of the 80 predictors is reported in Appendix A.

b) We defined 6 different outcomes for our analysis: Hypoglycaemic Events, Ketoacidosis, Nephropathy, Neurobehavioral complications, Neuropathy and Retinopathy. For each different outcome, a separate model has been derived by employing the 80 baseline predictors and the experimentation protocol explained in Section 4.2.3. The outcomes definition is explained in detail in the next sub–sections[1].

Hypoglycaemic Events

Hypoglycaemia consists in an abnormal decrement of sugar in the blood. The main adverse effect of hypoglycaemic events lies in a sudden shortage of sugar for the brain, which can lead to serious consequences (e.g., coma or death). During the DCCT study, each patient was reporting her hypoglycaemic events at each visit, specifying date and intensity of the event. We considered only serious hypoglycaemic events, i.e. events that led to coma or seizure, while building the predictive model. The time to event is measured in days from the baseline visit until the first hypoglycaemic event.

[1] Hypoglycemia and Ketoacidosis are usually classified as “short term” complications. However, the median time-to-event for DCCT patients that experienced Hypoglycemia or Ketoacidosis is 2.4 and 2.8 years, respectively. Thus we decided to develop long term risk assessment models also for these two complications.

Ketoacidosis

Ketoacidosis is an adverse event that happens predominantly in patients with T1DM (even though it can occur to T2DM patients as well). Ketoacidosis results from a shortage of insulin; in response the body switches to burning fatty acids and producing acidic ketone bodies that cause most of the symptoms and complications. Ketoacidosis events were reported by the patients at every visit. The time to event is measured in days starting from the baseline until the first occurrence.

Nephropathy

The term nephropathy covers a large spectrum of pathological conditions; it refers to any disease or damage that affects kidney functionalities. We decided to focus our attention on the clinical condition known as "Microalbuminuria", i.e. the abnormal presence of the albumin protein in the urine. Microalbuminuria is clinically defined as a value of albuminuria above 30 mg/24h in a urine test over twenty four hours [UK Renal Association (2011)]. DCCT patients were performing a 24h urine test once a year; time to event is thus measured in years starting from the baseline visit.

Neurobehavioral complications

The neuropsychological condition of DCCT patients was evaluated through 8 different tests at baseline, at 2, 5, 7 years and at the end of the study [DCCT Group (1996)]. A computerized algorithm [Lan et al. (1994)] was employed for identifying patients that could have had a sensible deterioration from the baseline; such patients were then independently evaluated by two experts, in order to confirm the presence of a relevant worsening.

We define as neurobehavioral event any significant worsening, detected by *at least* one expert, in the neurobehavioral condition with respect to the baseline.

Neuropathy

Diabetes can seriously damage the nerves of the body, especially in older people. A complete clinical neurological examination was performed on each DCCT patient biennially. We defined as Neuropathy event any abnormal response in the Autonomous Nervous System, i.e., any of the following conditions:

1. R-R variation (mean resultant) < 15.0
2. R-R variation < 20.0 in combination with Valsalva ratio < 1.5
3. Orthostatic hypotension caused by autonomic neuropathy as indicated by a decrease of at least 10 mm Hg in diastolic blood pressure in postural studies confirmed by blunted norepinephrine response in plasma catecholamine specimens

Retinopathy

Retinopathy refers to any damage of the retina, inflammations excluded. Retina conditions were assessed every three months during the DCCT study, through the "Early Treatment Diabetic Retinopathy Study" scale (ETDRS, [ETDRS (1991)]). A Retinopathy event is defined as a three point worsening with respect to the baseline value in the ETDRS scale that persists for over six months [DCCT Group (1993)]. Note that the time to event is calculated from the baseline visit to the beginning of the six months.

4.2.3 Experimentation protocol

We adopted a combined Machine Learning and Statistical approach for inducting, evaluating, and selecting a final predictive model for the time-to-retinopathy from survival data. "Survival" data are data measuring the time-to-event of interest (historically death, hence the name "survival data"). Survival data, as in the DCCT study, are often censored, i.e., for a given patient in the data that did not experience the event of interest (retinopathy) we do not know exactly when they will develop the complication: the information is censored. All we know is they did not experience the event during the time of the study. Censoring requires special consideration and specialized methods that can deal with such data. We set forth the following objectives in our analysis:

1. Induce the most accurate model $S(t, x_1, \cdots, x_n)$.

This in turn implies that we need to consider and evaluate several methods for learning such models and thus,

2. Consider and evaluate all state-of-the-art survival analysis regression methods for inducing model *S*.

Given that the inference of $p(x_1, ..., x_n \mid evidence)$ and the number of terms in the sum of $S(t, evidence) = \sum_{x_1,\cdots,x_n} S(t, x_1, \cdots, x_n) \cdot p(x_1, \cdots, x_n \mid evidence)$ grows exponentially with the number of possible patterns for $x_1, ..., x_n$ it is desirable to identify the minimal-size, yet close-to-optimally predictive subset of variables $x_1, ..., x_n$ that enter the model. Presumably also, the most-predictive, minimal-sized variable subset provides intuition into the mechanisms causing the disease, results in models easier to verify, understand and visualize. Variables not included in this set are either irrelevant or redundant given the selected variables. Thus we strived to,

3. Identify the minimal-size, close-to-optimally predictive subsets of clinical parameters to enter the model.

Finally, it is not enough to construct such a model without also providing an estimate of its performance, so that its utility can be assessed.

4. Provide an unbiased estimate of the performance of the induced model.

To achieve all of the above four goals and produce a predictive model, perform variable selection, optimize the learning parameters of the algorithms, and provide an unbiased estimation of the performance of the final model we employ a computational-experimentation protocol known as the nested-cross validation. The exact methodology and details on the methods, additional references, etc. are also described in a recent publication [Lagani and Tsamardinos (2010)].

The following learning algorithms and methods are employed:

1. Survival Data Regression Functions (along with the parameters tried during the analysis):
    - α. Cox Regression
        - ι. No parameters
    - β. Ridge Cox Regression
        - ι. Thirteen different values of the shrinking parameter
    - χ. Accelerated Failure Models (AFT)
        - ι. Parametric distributions tried: Extreme, Logistic, Gaussian, Weibull, Exponential, Lognormal and Log-logistic
    - δ. Random Survival Forest (RSF)
        - ι. Varied number of splitting point per variable (1, 5, 10)
    - ε. Support Vector Machine Censored Regression
        - ι. Several kernels and cost parameter c within {1, 10, 100}

2. Variable selection methods:
    - α. No Selection
        - ι. No parameters
    - β. Univariate Association Selection
        - ι. Using the log-rank test
        - ιι. P-values threshold {0.05 , 0.1}
    - χ. Forward Step-wise Selection
        - ι. Using a type of log-rank test

ιι. Percentage of variables to be selected {1%, 5% , 10%}

δ. Lasso regression

ι. Shrinking parameter {0.1, 1, 10}

ε. Bayesian Variable Selection (BVS)

ι. Hyperparameters corresponding to standard deviation of the priors {0.1, 0.5, 1}

ϕ. Survival Max Min Parents and Children (SMMPC)

ι. max-conditioning set k in {2, 3}, p-value threshold in {0.05, 0.1}

3. Performance metrics:

α. Concordance Index

ι. Similar to Area Under the Receiving Operating Characteristic Curve (AUC) for survival data

ιι. Interpretation: given two random patients, the CI is the probability of ranking their risk in the same order as the actual occurrence of the event to the patients

ιιι. 0.5 is the expected CI by chance

β. Integrated Brier Score

ι. Overall measure of the predictions *over time*

ιι. Compare the estimated survival functions with the survival information of the patients.

4. Parameter Optimization and Performance Estimation

α. Nested 10-fold cross-validation scheme is used to *avoid multiple testing problems* in estimation of performance; extends training-validation-test scheme to cross-validation.

ι. The inner cross-validation is employed for selecting the best combination of methods and parameters; it returns a single model for each test-set of the outer loop

ιι. The outer cross-validation is employed for estimating the performance expected by the learning method that includes selection of the best parameters

ιιι. Final model is built on all data

β. Determine statistical significance of the differences between the models produced by the algorithms

ι. Permutation-based test (1000 permutations) to estimate the empirical distribution of the differences under the null assumption of equal CIs

This experimentation protocol leads to the evaluation of approximately 43000 models; the nested cross validation procedure is adopted specifically for avoiding the possibility of overfitting. It has been employed in several high-visibility publications [Aliferis et al. (2010a)], [Aliferis et al. (2010b)], [Statnikov et al. (2005)], [Aliferis et al. (2009)].

4.3 New Knowledge gained from available data set - Results

Tables in the following subsections show the nested – cross validated Concordance Index (CI) results for each outcome and for each combination of regression algorithms and feature selection methods. Values in parenthesis indicate the number of variables averagely selected. We employed a permutation based statistical test in order to assess whether the difference in terms of CI between

SMMPC and any other feature selection method is statistically significant (p–value < 0.05). Methods that perform statistically better than SMMPC are marked in red, while methods with poorer performances are in blue. Correspondent Integrated Brier Score values are reported in Appendix B.

For each outcome we selected a single combination of feature selection methods and regression algorithms as the "best combination". The criteria for the selection are (in order of importance):

1. Predictive power
2. Parsimony of the model (the lesser variables, the better)
3. Understandability of the model

In few cases we preferred to discard the most well performing solution in favour of another highly performing model with fewer variables. The "best combination" for each outcome is underlined in the corresponding table.

After selecting the best combination of feature selection method and regression algorithm for each outcome, we employed a cross validation procedure for determining the best parameters configuration, i.e., the parameters configuration corresponding to the maximal cross validated CI. Finally, for each outcome we trained a single model on its respective, whole dataset. The next subsections report, for each outcome, the characteristic of the respective best model.

Hypoglycaemia outcome

For the severe hypoglycaemia outcome the best combination of feature selection and regression algorithm is {SMMPC + SVCR} (see following table).

	Cox Regression	AFT	RSF	Ridge Cox Regression	SVCR
No Selection	0.63 (100.0)	0.63 (100.0)	0.67 (100.0)	0.64 (100.0)	0.54 (100.0)
Univariate Selection	0.65 (25.1)	0.65 (25.1)	0.66 (25.2)	0.66 (25.1)	0.61 (23.9)
Forward Selection	0.63 (26.0)	0.63 (17.2)	0.66 (64.4)	0.65 (12.0)	0.65 (12.0)
Lasso Selection	0.54 (10.0)	0.54 (10.0)	0.54 (10.0)	0.52 (10.0)	0.51 (10.0)
BVS	0.65 (24.1)	0.65 (25.0)	0.67 (27.4)	0.66 (25.8)	0.60 (21.9)
SMMPC	0.66 (6.5)	0.66 (6.6)	0.65 (6.2)	0.66 (6.1)	**0.67 (5.7)**

Table 1: CI results for the Hypoglycaemia outcome

The cross validation procedure selected the following, optimal parameters configuration:

- SMMPC:
  - Significance threshold: 0.05
  - Maximal size of conditioning set: 2
- SVCR:
  - Linear kernel
  - Cost parameter C: 10

The final selected variables, along with their respective weights in the linear Support Vector Machine, are the following:

Variables	Description	Linear Weight
GROUP	Randomization Group (1 = strict control, 0= standard)	-384.99
INSULIN	Total Insulin Dosage Units/Weight (kg)	-133.85
PRIORHYP	Past History of Severe Hypo	-143.84
FAMNIDDM	Family History of NIDDM (yes/no)	120
BCVAL5	Stimulated C-Peptide (pmol/ml)	-5.07
OBMARRY=2	Marital Status (1 = married, 0 = no/divorced)	80

Table 2: most predictive variables for the hypoglycaemia outcome, along with their linear SVM weights

The Support Vector Machine for Censored Target attempts to directly predict the time to event; this means that negative coefficients imply a higher probability of experiencing a severe hypoglycaemia while positive coefficients are associated with a lower probability of hypoglycaemia. Thus, the model suggests that:

Following a stricter protocol of blood glucose control actually increases the probability of hypoglycaemia (as pointed out in [DCCT Group (1995)]).

1. High dosages of insulin, past history of hypoglycaemic events, high values of stimulated C – Peptide, are all associated with higher probability of severe hypoglycaemia
2. A family history of not-insulin dependent diabetes mellitus (NIDDM) and being married is associated with a lower probability of experiencing a hypoglycaemic event with hospitalization or coma.

Ketoacidosis outcome

The nested cross validated CI results are reported in the following table; in all cases, a relatively elevated number of features is needed in order to achieve good performances. We selected the combination {Univariate Selection + Ridge Cox Regression} as best solution.

	Cox Regression	AFT	RSF	Ridge Cox Regression	SVCR
No Selection	0.69 (100.0)	0.68 (100.0)	0.65 (100.0)	0.70 (100.0)	0.60 (100.0)
Univariate Selection	0.68 (41.6)	0.68 (42.3)	0.65 (43.5)	**0.70 (44.5)**	0.62 (40.2)
Forward Selection	0.68 (45.8)	0.69 (50.2)	0.63 (54.0)	0.69 (56.4)	0.59 (34.9)
Lasso Selection	0.63 (10.0)	0.62 (10.0)	0.61 (10.0)	0.62 (10.0)	0.57 (10.0)
BVS	0.68 (29.2)	0.68 (29.6)	0.64 (32.9)	0.69 (31.1)	0.64 (24.5)
SMMPC	0.62 (11.0)	0.63 (11.3)	0.62 (10.7)	0.63 (11.1)	0.64 (11.2)

Table 3: CI results for the Ketoacidosis outcome

The optimal parameters configuration chosen by the single cross validation procedure is the following:

- Univariate Selection:
  - Significance threshold: 0.05
- Ridge Cox Regression:
  - Shrinkage parameter: 22.52

We report the 5 parameters with the highest coefficient (in absolute value) in the Ridge Cox regression Model. Positive coefficients indicate higher risk of developing ketoacidosis, and conversely for negative coefficients. Thus, high values of stimulated C – Peptide, as well as low values of cholesterol and T30 blood glucose profile are associated with low risk of ketoacidosis. Interestingly, previous history of ketoacidosis and belonging to a “poor” social class (as scored by the Hollingshead scale) are associated with higher risk of ketoacidosis.

Variables	Description	Coefficients
BCVAL5	Stimulated C-Peptide (pmol/ml)	-0.2452
BCVAL30A	T30-Blood Glucose Profile 6 mg/dl	0.1788
OBDKAHSP	Hospitalizations for DKA in Past Year	0.1594
CHOL	Cholesterol (serum,mg/dl)	0.1520
HOLLSCOR	Hollingshead (2 Factor) Social Class score	0.1396

Table 4: most predictive variables for the ketoacidosis outcome, along with their ridge cox regression coefficients

Nephropathy outcome

The best combination for the Nephropathy outcome consists in {SMMPC + Random Survival Forest}, with a nested cross validated CI of 0.69.

	Cox Regression	AFT	RSF	Ridge Cox Regression	SVCR
No Selection	0.65 (100.0)	0.65 (100.0)	0.69 (100.0)	0.69 (100.0)	0.54 (100.0)
Univariate Selection	0.68 (39.2)	0.68 (39.2)	0.69 (39.8)	0.69 (40.9)	0.60 (38.9)
Forward Selection	0.68 (15.6)	0.68 (15.6)	0.69 (39.6)	0.68 (35.0)	-
Lasso Selection	0.60 (10.0)	0.60 (10.0)	0.58 (10.0)	0.60 (10.0)	-
BVS	0.69 (18.1)	0.68 (18.7)	0.69 (21.9)	0.69 (19.0)	-
SMMPC	0.69 (8.4)	0.69 (8.3)	**0.69 (8.1)**	0.69 (8.6)	0.6787

Table 5: CI results for the Nephropathy outcome

The optimal parameters configuration identified by the single cross validation procedure is the following:

- SMMPC:
  - Significance threshold: 0.05
  - Maximal size of conditioning set: 3
- RSF:
  - Number of splitting points per variables: 5

The following table reports the variables involved in the final models along with their relative importance, as calculated by the randomSurvivalForest R package [Ishwaran and Kogalur (2007)]. The variable with relative importance equal to 1 is the most relevant for the classification; variables with relative importance of 0 do not provide any significant contribution to the model.

Variables	Description	Relative importance
AER	Albuminuria (mg/24hr)	1.00
AGE	Age at baseline	0.5213
GROUP	Randomization Group (1 = strict control, 0= standard)	0.2180
RETPAT00	Retinopathy Severity Level	0.1890
OBWEIGHT	Weight (kg) at Baseline	0.0808
HBA00	Hemoglobin A1c at Baseline	0.0181

Table 6: most predictive variables for the Nephropathy outcome, along with their relative importance

Neurobehavioral outcome

The predictive performance for the Neurobehavioral outcome is generally quite poor (i.e., below 0.6; notice that CI = 0.5 corresponds to the performance expected from the random classifier). The combination with the highest CI value is {Bayesian Variable Selection + SVCR}.

	Cox Regression	AFT	RSF	Ridge Cox Regression	SVCR
No Selection	0.53 (100.0)	0.53 (100.0)	0.54 (100.0)	0.53 (100.0)	0.57 (100.0)
Univariate Selection	0.55 (19.3)	0.56 (20.3)	0.52 (18.9)	0.51 (19.0)	0.54 (20.6)
Forward Selection	0.52 (24.2)	0.55 (29.4)	0.53 (36.7)	0.50 (30.8)	-
Lasso Selection	0.46 (10.0)	0.45 (10.0)	0.46 (10.0)	0.42 (10.0)	-
BVS	0.54 (29.6)	0.57 (29.6)	0.53 (31.7)	0.52 (29.2)	**0.58 (30.7)**
SMMPC	0.55 (5.6)	0.56 (5.6)	0.55 (5.4)	0.54 (5.3)	0.53 (5.8)

Table 7: CI results for the Neurobehavioral outcome

The cross validation procedure selected the following, optimal parameters configuration:

- BVS:

- Hyper-prior on the parameters: 0.5
- SVCR:
  - Polynomial kernel with degree 2
  - Cost parameter C: 1

Neuropathy outcome

The combination {SMMPC + Random Survival Forest} achieves a nested cross validated performance CI = 0.79 with only ~5 variables (on average across the external folders).

	Cox Regression	AFT	RSF	Ridge Cox Regression	SVCR
No Selection	0.74 (100.0)	0.71 (100.0)	0.75 (100.0)	0.74 (100.0)	0.59 (100.0)
Univariate Selection	0.79 (22.3)	0.79 (22.4)	0.78 (20.0)	0.76 (21.1)	0.62 (21.7)
Forward Selection	0.77 (11.6)	0.77 (11.6)	0.77 (20.6)	0.76 (20.2)	0.71 (12.0)
Lasso Selection	0.66 (10.0)	0.63 (10.0)	0.60 (10.0)	0.63 (10.0)	-
BVS	0.76 (17.1)	0.76 (16.8)	0.76 (19.5)	0.76 (19.5)	-
SMMPC	0.77 (5.5)	0.78 (5.4)	**0.79 (4.7)**	0.77 (5.6)	0.7214

Table 8: CI results for the Neuropathy outcome

The optimal parameters configuration identified by the single cross validation procedure is the following:

- SMMPC:
  - Significance threshold: 0.1
  - Maximal size of conditioning set: 3
- RSF:
  - Number of splitting points per variables: 10

The following table reports the variables involved in the final models along with their relative importance.

Variables	Description	Relative importance
RRV00	ANS - RR Variation (x 1000)	1.00
RETPAT00	Retinopathy Severity Level	0.2284
AER	Albuminuria (mg/24hr)	0.0713
HBA00	Hemoglobin A1c at Baseline	0.0287
F002DATE=87	Baseline visit performed in 1987	0
OBPATJOB=4	Patient's Occupation (1 = clerical or similar, 0 = other)	0

Table 9: most predictive variables for the Neuropathy outcome, along with their relative importance

The most relevant variable for a correct risk evaluation is the RR variation, i.e., a measure of the status of the Autonomous Nervous System (ANS). Others relevant variables are the Retinopathy Severity level, Albuminuria and the HA1c, all measured at baseline. It is interesting to note that the random survival forest model discards as useless the variables “F002DATE=87” and “OBPATJOB=4”, which were previously selected as relevant by the SMMPC algorithm.

Retinopathy outcome

We chose the combination {SMMPC + Ridge Cox Regression} as best configuration for the Retinopathy outcome. Even though this configuration does not provide the best nested cross validated CI, it is a convenient compromise among good performances, model parsimony and interpretability of the model.

	Cox Regression	AFT	RSF	Ridge Cox Regression	SVCR
No Selection	0.70 (100.0)	0.68 (100.0)	0.75 (100.0)	0.72 (100.0)	0.62 (100.0)
Univariate Selection	0.73 (32.7)	0.73 (32.7)	0.75 (32.7)	0.73 (32.7)	0.71 (32.7)
Forward Selection	0.70 (64.0)	0.71 (64.0)	0.74 (64.0)	0.73 (64.0)	0.64 (64.0)
Lasso Selection	0.59 (10.0)	0.59 (10.0)	0.58 (10.0)	0.59 (10.0)	0.58 (10.0)
BVS	0.73 (14.3)	0.74 (14.3)	0.73 (14.3)	0.74 (14.3)	0.71 (14.3)
SMMPC	0.73 (4.0)	0.74 (4.0)	0.75 (4.0)	**0.74 (4.0)**	0.70 (4.0)

Table 10: CI results for the Retinopathy outcome

The cross validation procedure selected the following parameters configuration as optimal:

- SMMPC:
  - Significance threshold: 0.05
  - Maximal size of conditioning set: 2
- Ridge Cox Regression:
  - Shrinking parameter: 21.8750

The following table reports the variables employed in the final model along with their coefficients. Positive coefficients indicate higher risk of developing Retinopathy, while negative coefficients denote a low risk of experiencing the adverse event. Thus, the model suggests that the experimental treatment improves the prognosis, while a high concentration of glycated haemoglobin and an already severe Retinopathy at baseline will probably denote a poor outcome.

Variables	Description	Coefficients
GROUP	Randomization Group (1 = strict control, 0= standard)	-0.71
HBA00	Hemoglobin A1c at Baseline	0.48
RETPAT00	Retinopathy Severity Level	0.25

Table 11: most predictive variables for the Retinopathy outcome, along with their ridge cox regression coefficients

4.3.1 A structural approach to the analysis of the DCCT data

We employed the algorithm Max – Min Hill Climbing, described in [Tsamardinos et al. (2006)], for reconstructing the structure of the Bayesian network representing the DCCT data distribution. The resulting Direct Acyclic Graph (DAG) is shown in Figure 1; violet nodes represent measurements at baseline, while the six outcomes are in green.

We plan to employ the Bayesian Network trained on DCCT baseline visit data for allowing the Long Term Risk Assessment models providing predictions even when some of the relevant parameters are missing. Let us represent the parameters employed by a specific model with the variables $x_1 \ldots x_n$. Let us also define the conditional joint distribution $p(x_1, \ldots, x_n \mid evidence)$, where *evidence* is a vector with the values of any subset of the clinical parameters in **Fehler! Verweisquelle konnte nicht gefunden werden.**. When some of the input variables $x_1 \ldots x_n$ are missing, we can then compute the survival function to return as:

$$S(t, evidence) = \sum_{x_1, \cdots, x_n} S(t, x_1, \cdots, x_n) \cdot p(x_1, \cdots, x_n \mid evidence)$$

Given the network (structure and the parameters of the network) the joint $p(x_1, \ldots, x_n \mid evidence)$ can be computed by using standard network inference algorithms (e.g., the Junction-Tree algorithm [Lauritzen and Spiegelhalter (1988)]).

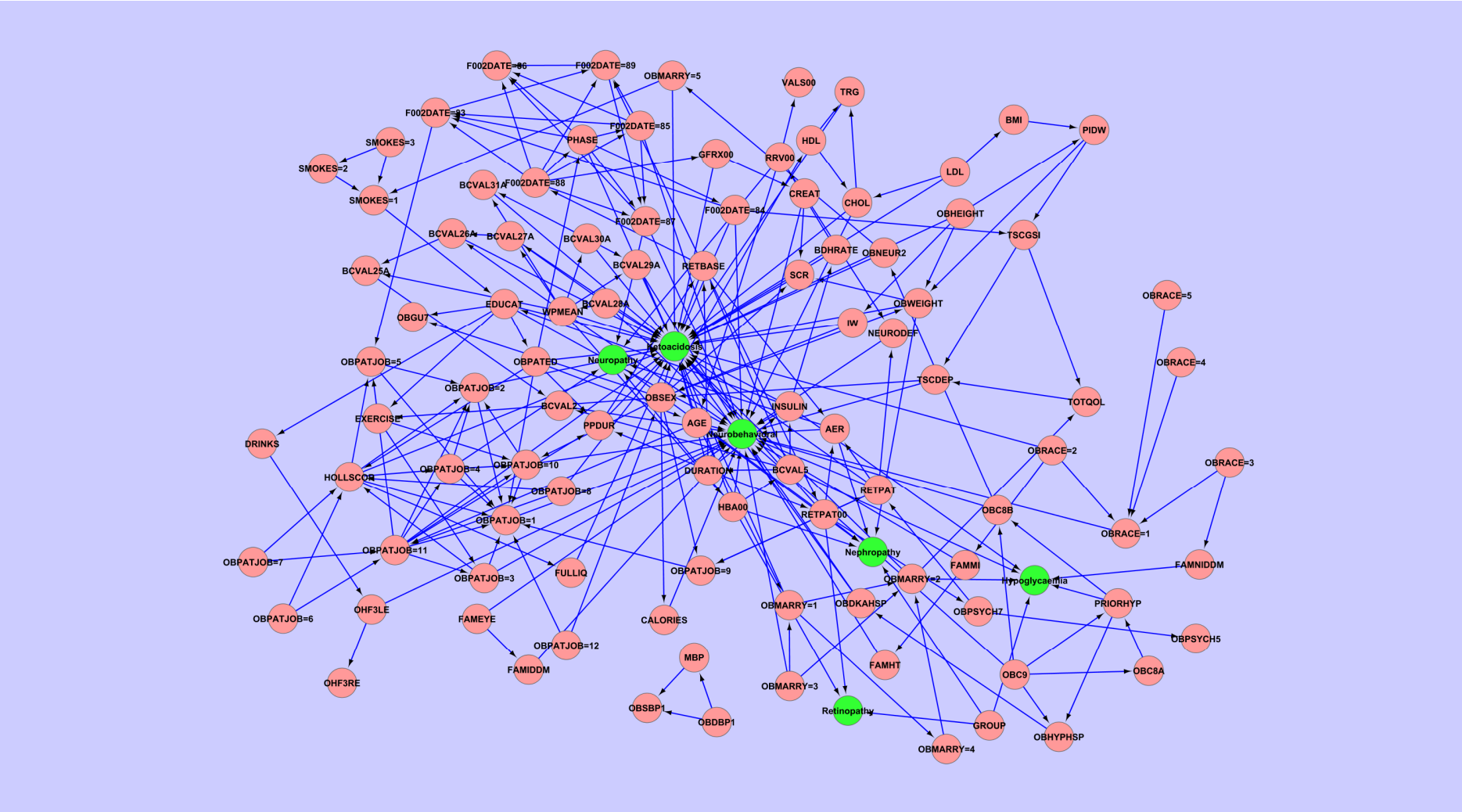

Figure 1: Direct Acyclic Graph representing the DCCT data distribution

4.4 Limitations

There are two major factors that may undermine the generality and applicability of our results: (a) time and location of DCCT data collection, and (b) the experimental design of the DCCT study.

a) The DCCT data were collected in a given geographical area (USA and Canada), and in a period of time spanning from 1983 to 1993. In the best of possible cases, one may assume that the DCCT data provide a faithful "picture" of the North America diabetes patients' community towards the end of the previous century. However, **there is no indication whether the results obtained from the DCCT data may be valid nowadays for populations residing in other geographical regions.** Even though it is reasonable to assume that diabetes pathophysiology mechanisms did not significantly change in the last decades, there are several important factors that are undoubtedly different across time and space. One of the most relevant of such factors is medication: medical procedures underwent several significant changes in the last thirty years, and hopefully these changes represent improvements. Thus any assessment of individual risk performed on the basis of the DCCT data may be seen as a worst-case-scenario evaluation. Another factor undergoing major changes depending on time and location is nutrition: North America diet is, on average, extremely diverse from the nutritional regimen of south Italy inhabitants (for example).

   From a statistical point of view, we can say that the distribution of the data collected on the DCCT sample may drastically differ from the distribution of the same variables in other populations. Thus, any prediction based on risk assessment models derived from the DCCT data should be interpreted having in mind the possible difference between the DCCT population and the reference population of the patient.

b) The DCCT study consisted in a randomized case control study, comparing two different treatment regimens (standard versus tight glucose control regimens). This implies that a risk assessment evaluation performed on the basis of the DCCT data may be meaningful only if the patient follows one of these two treatments, or, at least, a similar one.

   A more relevant, but also more subtle implication of the DCCT study design is the following: within each experimental group, DCCT data are purely observational. This means that **it is not possible to predict the effect of an intervention on the basis of DCCT data, except if the intervention consists in switching from a treatment regimen to the other one**. For example, the results provided by the Retinopathy model indicate that the higher the concentration of glycated haemoglobin (HA1c), the worst the prognosis. However, adopting any measure that will result in a reduction of this clinical parameter will not ensure an improvement of the prognosis. High concentration of glycated haemoglobin and Retinopathy adverse events may be associated by means of a third variable that causes both, and thus intervening on the HA1c value does not affect the prognosis. Or, in a somehow more threatening case, the measures adopted for diminishing the concentration of HA1c may also increase the probability of developing a Retinopathy, leading to the outcome one wanted to avoid. On the other hand, our analysis confirmed that patients following the standard treatment regimen experienced more often Retinopathy than the patients belonging to the other group. Thus, it is possible to assert that forcing a patient to switch from a normal treatment to the regimen with stricter glucose control will improve his probability of avoiding Retinopathy. However, it must be clear that such conclusion is possible only because DCCT data were randomly assigned between the two treatment regimens.

4.5 Outlook

Future work for the long term risk assessment models will mainly follow two directions:

1. In-deep interpretation of the obtained results in the context of the current medical knowledge. An impressive number of works has been published on diabetes complication risk factors. Our results on the DCCT dataset need to be contrasted against the most relevant and reliable findings present in the literature.
2. Devising and evaluating long term risk assessment models for other complications. A large number of clinical parameters was measured during the DCCT study. Thus, several negative outcomes can be defined on the basis of the DCCT data. We will investigate the possibility of defining and analysing new negative events, with an equal of greater relevancy, from the clinical perspective, with respect to the six ones already defined and analysed.

5. Physiology-based glucose-insulin model

5.1 State-of-the-art and related work

State-of-the-art model-based inference/risk analysis from measurement (time-series) data during glycaemic control focuses on two aspects:

1. Safety measures using hypoglycaemia-detection algorithms to prevent hypoglycaemic episodes, mainly postprandial and nocturnal.
2. Estimation of insulin resistance and related pathophysiological conditions (especially in intensive care unit (ICU) environments)

The physiology-based glucose-insulin model has been developed to allow the adaption of similar methods to evaluate pathophysiologic conditions of patients. However, for an operative and validated realization, extensive time-series datasets of blood glucose levels from patients with the conditions of interest would be necessary. To date, these datasets are not available.

In the following, some background information and related work of the two main aspects of model based inference from time-series data is given.

5.1.1 Hypoglycaemia detection

In current research on simulation- and model-based glucoregulatory systems (artificial pancreas systems), no efforts on model-based hypoglycaemia prediction using covariates other than glucose have been made. Continuously measured blood glucose levels are the current state-of-the-art for the prediction of hypoglycaemic episodes as published by (Palerm, Willis et al. 2005; Cameron, Niemeyer et al. 2008; Buckingham, Chase et al. 2010; Dassau, Cameron et al. 2010; Eren-Oruklu, Cinar et al. 2010).

The physiology-based model kernel which is developed within REACTION describes the detailed pharmacokinetics-dynamics and molecular interactions of glucose, insulin and glucagon similar but in more detail than other current glucoregulatory models by (Hovorka, Chassin et al. 2004; Dalla Man, Rizza et al. 2007). The immediate glucoregulatory model is not yet developed for a multi-parametric approach to hypoglycaemia detection with the integration and prediction of parameters such as heart rate, heart rate variability and QT intervals, galvanic skin response or skin temperature.

The interrelation of heart rate (variability), blood flow, blood pressure and oxygen demand will be investigated. For this, the development of a cardiovascular systems model is being investigated. As an interface between the glucoregulatory and the cardiovascular model, the contribution of exercise to blood glucose variability in relation to heart rate and oxygen consumption is also evaluated. However, mechanisms governing the cardiovascular system based on nutritional demand are complex and poorly understood and have, to the best of our knowledge, never been modelled before. Thus, modelling efforts that could describe the interrelation of glucose levels and heart rate remain scientifically challenging.

Multiparametric sensor measurements are expected to be used as additional indicators during predicted hypoglycaemic episodes to increase the specificity of predicted low blood glucose levels. Thus, multiparametric sensors in combination with a simulation model should offer a better means to predict low blood glucose levels than either technique alone.

5.1.2 Insulin resistance and related pathophysiological conditions

The term “Insulin Resistance” describes a reduction of “Insulin Sensitivity” and is usually used to refer to the defective regulation of carbohydrate metabolism by insulin. Insulin sensitivity is characterized as the “gearing” ratio for insulin-dependent glucose uptake in muscle and adipose tissue. This disorder has been quantified by numerous methods (Kumar and O'Rahilly 2005), usually by measuring the amount of glucose infused to maintain euglycemia at a fixed insulin concentration (glucose clamp). The main consequences of insulin resistance include (Andrews and Walker 1999):

- Impaired insulin-dependent down-regulation of hepatic glucose release.
- Impaired insulin-mediated increase in peripheral glucose uptake.

A reduction in insulin sensitivity, called a state of insulin resistance, therefore impairs the body's ability to control its blood glucose levels and reduce hyperglycemia. The causes for the reduction in insulin sensitivity are many-sided, making insulin sensitivity a multi-parametric property.

In general, the interaction of signalling proteins, including insulin itself, with the insulin signalling pathway are thought to cause these changes in insulin sensitivity. Known mechanisms to reduce insulin sensitivity are reduction in receptor transcription and receptor translocation or effects on downstream signalling like serine phosphorylation of insulin receptor substrate-1 (IRS-1) or activation of protein kinase C (PKC). Several of these crosstalk signalling pathways are known (Kumar and O'Rahilly 2005).

Insulin sensitivity is of major importance at intensive care units, where the effects of inflammation and its treatment, e.g. with anti-inflammatory steroids like cortisone, can cause severe dynamic changes in insulin sensitivity.

Model-based approaches to assess insulin resistance include work by Chase et al. who are working on the prediction of sepsis using on model-based estimates of insulin sensitivity derived from blood glucose measurements (Blakemore, Wang et al. 2008; Lin, Parente et al. 2011). A parsimonious model based on Bergman's Minimal Model is used for data analysis (Hann, Chase et al. 2005) and estimation of insulin sensitivity (Lin, Lee et al. 2008).

Hovorka et al. also estimate changes in insulin sensitivity during blood glucose control in an ICU setting (Hovorka, Chassin et al. 2008). However, they have not described if or how the estimated insulin sensitivity values could be used for further analysis.

Currently, no algorithms have been implemented to detect hypoglycaemia based on measurement data. However, the predictive power of the physiology-based models can be used to predict future trends of blood glucose in detail.

Also, using the physiology-based glucose-insulin model, a similar approach to evaluate pathophysiologic conditions of patients from the multi-parametric properties of insulin resistance could be adapted. However, for this, extensive time-series datasets of blood glucose levels from patients with the conditions of interest would be necessary. To date, these datasets are not available.

The development of the physiology-based model of the glucose-insulin metabolism is documented in detail in deliverable D6-4-1. A validation of the model in a clinical context is planned. Beyond this effort, there are no efforts yet for additional model-based data analysis for risk assessment.

5.2 Methods and tools in physiology-based pharmacokinetic / pharmacodynamic modelling

In the pharmaceutical field, an iterative process is established in which generic PBPK models are informed by experimental data and subsequently used to predict outside of the experimental data space. Using this step-wise procedure, all experimental data can be effectively integrated into the PBPK model as it becomes available over time, as is the case in clinical development.

Here, PBPK models are developed on a restricted cohort of adults that participated in glucoregulatory-related clinical trials as well as on Data from standard tolerance tests (glucose and insulin) and clamp studies. The data from these clinical trials was used for both, model development and validation, as each patient underwent two independent trials. Thus one dataset per patient was used for model development, while the second set was then used for model validation. Once this adult PBPK model has been established, it can be extrapolated to individuals, be it adults, children or elderly outside the selected cohort.

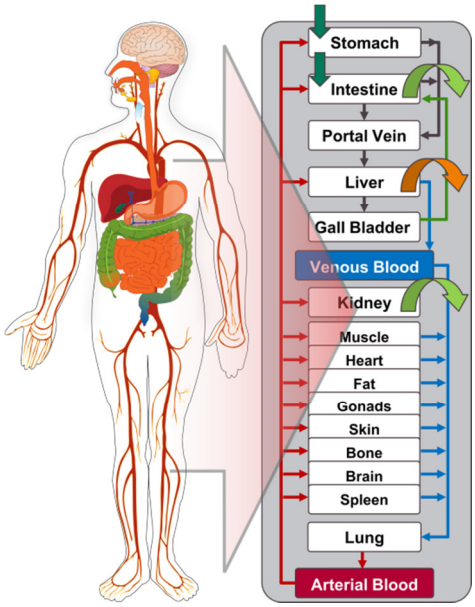

Figure 2: PBPK model structure implemented in PK-Sim®.[2]

This is done by incorporation of a priori knowledge into the model that is relevant for predicting the pharmacokinetics of a drug in adults with varying anthropometric properties. Required knowledge on anthropometric properties includes the anatomical and physiological changes associated with age, weight, height, gender and race (e.g. age dependence of body weights, organ weights, blood flows, etc.) as well as an understanding of how the activity of active processes (e.g. clearance processes, transporters), which affect the drugs pharmacokinetics, scale with relation to these properties. The process of transferring physiological properties into am model is depicted as a schematic in Figure 2.

These resulting models allow dosing recommendations or the analysis of the effects of alternative administration scenarios or physiological changes associated with disease on drug pharmacokinetics- and dynamics.

The PBPK models for glucose, insulin and glucagon have been established in PK-Sim® 4.2 and coupled in MoBi® 2.3. Model identifications and model parameterization have been conducted using the MoBi® Toolbox for MATLAB® 2.2 (all tools have been developed by Bayer Technology Services GmbH, Germany, www.systems-biology.com; MATLAB® is a product of The MathWorks Inc., USA). A general description of the software platform including a detailed example of how to build parent-metabolite models and simulate populations with polymorphic enzymes has been published recently (Eissing, Kuepfer et al. 2011). The software settings and parameter sets needed for the model developed here are detailed in D6-4-1. In the following two sections a brief overview on the modelling platforms PK-Sim® 4.2 and MoBi® 2.3 is given.

5.2.1 PKSim®

PK-Sim® represents a commercial software package for PBPK modelling, developed by Bayer Technology Services GmbH. All important absorption, distribution, metabolization and excretion (ADME) processes are implemented. PK-Sim® can be used generically because the model structure is independent from the substance to be examined, i.e. without modification of the underlying model structure, compounds can be defined and compound parameters can be modified (Bayer-Technology-Services 2009). PK-Sim® enables the simulation of drug distribution dynamics resulting from single and multiple applications of drugs, through all relevant routes of administration, which, if necessary can also be customized in MoBi®.

PKSim® has a number of different basic model structures available for choice. Based on the biochemical characteristics of the respective drug, other aspects of the fluid circulation, e.g. lymph flow for large proteins, become more important. Each compartment, i.e. organ as illustrated in Figure 2, is

[2] The basic model structure can be extended by additional processes such as active transporters or enzyme-mediated metabolization (Eissing, Kuepfer et al. 2011)

generally divided into four sub-compartments as shown in Figure 3: intracellular space, interstitial space, blood cells and blood plasma. This model type was used for the simulation of glucose.

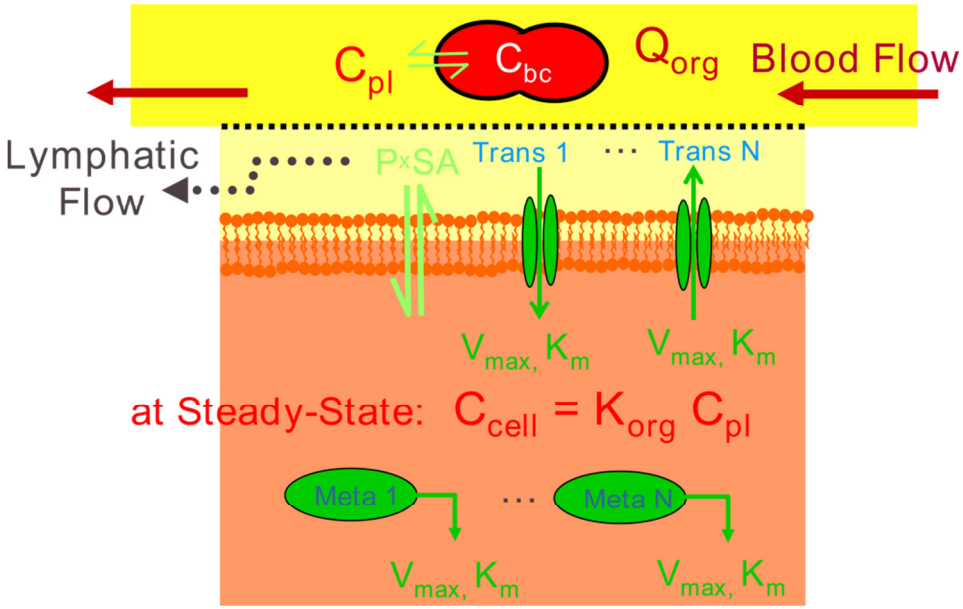

Figure 3: Organ representation in PK-Sim®[3]

For larger proteins like insulin and glucagon, the “two pore” model type can be chosen. Here the exchange across capillary walls is described additionally by the two pore theory, assuming convection and diffusion through two types of pores (Rippe and Haraldsson 1994).

The compound properties and the physiological properties of the subject define the passive distribution of the modelled substance within the organism. A detailed description of the equations can be found in (Eissing, Kuepfer et al. 2011).

5.2.2 MoBi®

The MoBi® software package from Bayer Technology Services GmbH is a tool for mechanistic and dynamic modelling of biological processes. It is fully compatible with PK-Sim®. Models that were created in PK-Sim® can be exported to MoBi® to be modified and/or coupled. This enables the development of complex PBPK models of a parent drug and its numerous metabolites or of models with interacting compounds as in the glucose metabolism with the interactions of glucose, insulin and glucagon, by connecting PK-Sim® models in MoBi®. In this work, all models created with PK-Sim® are exported to MoBi® for coupling via the known pharmacodynamic interactions (as described in D6-4-1), and for parameter identification.

[3] …including three sub-compartments (intracellular, interstitial, plasma and blood-cells) and the corresponding flow rates and transports. C: concentrations, Q: flow rates, P × SA: permeability surface area products, K: partition coefficient, Vmax, Km: Michaelis-Menten constants. From Willmann et al, 2003 (Willmann, Lippert et al. 2003).

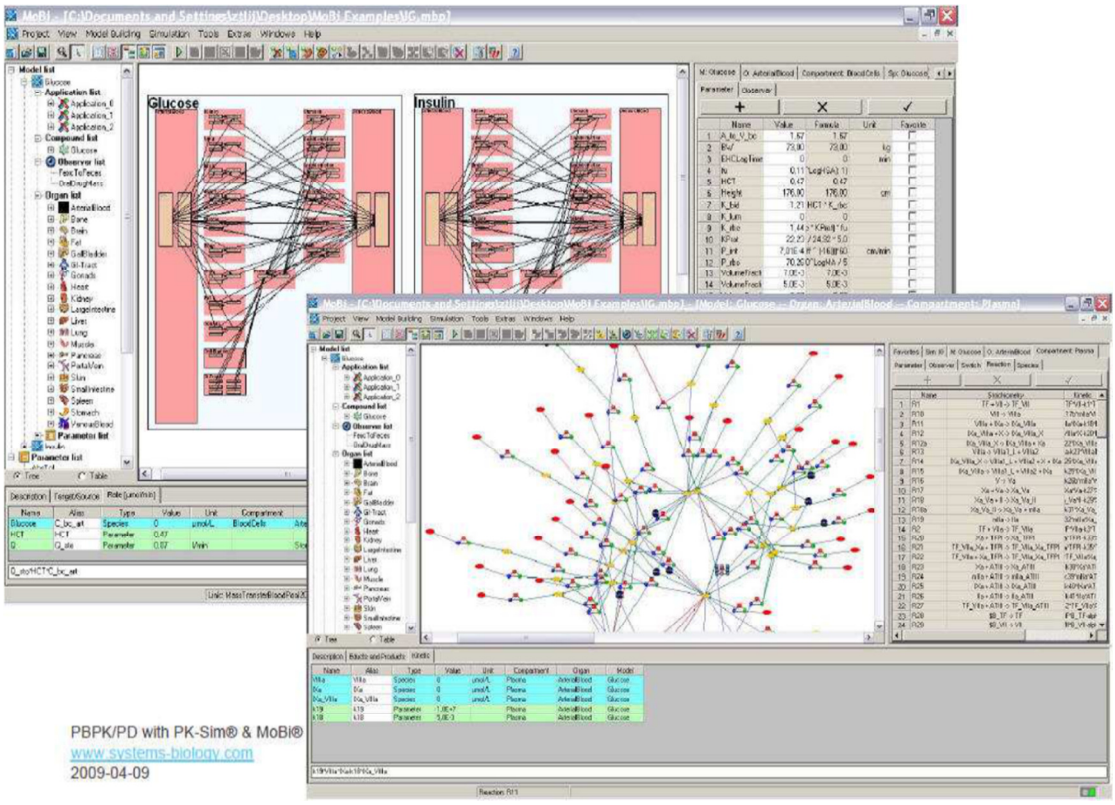

Figure 4: GUI of MoBi®

The Mobi® GUI allows the implementation of protein-protein interactions, clearance processes etc., as shown in Figure 4. Furthermore, by the integration of Matlab® via the MoBi® -Toolbox for Matlab®, complex optimizations and sensitivity analysis can be performed from which the results are presented in D6-4-1.

5.3 New Knowledge gained from available data set - Results

Besides data from literature, BTS has been working with a dataset from the REACTION partner MUG, which was generated during a clinical trial. In this trial, two algorithms, a CSII protocol and MPC control, were compared. 12 voluntary subjects with type 2 diabetes did a full day trial with overnight stay for each algorithm resulting in 24 whole-day datasets.

Two major new insights have been gained from our work with the dataset from MUG:

1. Clearance of insulin may be temporarily increased during a certain time period after a large insulin dose (e.g. an insulin bolus shot) thus suggesting dynamic changes in the specific rate of clearance
2. Insulin clearance is not only nonlinear for high insulin doses (saturation) but is over-proportionally high at low insulin concentrations (positive and negative cooperativity)

Saturation of insulin clearance is accounted for in a number of state of the art glucose-metabolism models (Hovorka, Chassin et al. 2008; Lin, Razak et al. 2011), however, none of the current state-of-the-art models used for blood glucose control (Dalla Man, Rizza et al. 2007; Hovorka, Chassin et al. 2008; Lin, Razak et al. 2011) or model-based inference (Dalla Man, Caumo et al. 2004) accounts for positive and negative cooperativity at low insulin concentrations or dynamic changes in insulin clearance. Although these phenomena have been the subject of research in the past (Wanant and Quon 2000), they have never been considered in the development of models of the glucose-metabolism. In the following, a short overview on the modelling methods and tools is given as well as the impact of cooperativity and dynamic insulin clearance will be elucidated.

5.3.1 Cooperativity

While working with the MUG dataset, it was observed, that at low insulin concentrations, caused by very low infusion rates in the CSII protocol or during pump shutdowns in the MPC protocol, there was a dramatic dropdown in insulin concentrations atypical for proportional insulin clearance under the assumption that binding affinity to the insulin receptor is constant. These observations are depicted in Figure 5 and Figure 6. The measured dropdown of plasma insulin levels suggests, when compared to the simulated concentration profile, that the current model assumptions are not sufficient at low insulin concentrations and, if this observation proves to be reproducible in further patients, need to be revised.

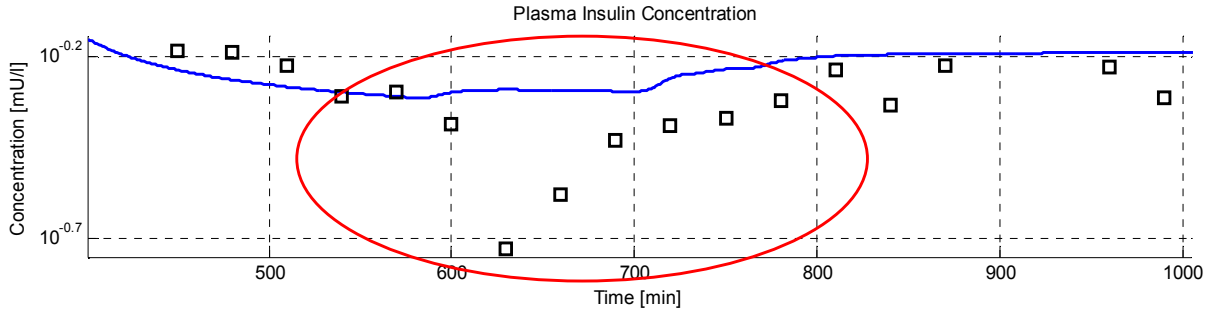

Figure 5: Simulated (straight blue line) and measured (boxes) plasma insulin concentration of Subject 6 (CSII protocol). The time-period of increased insulin clearance is marked with a red ellipse.

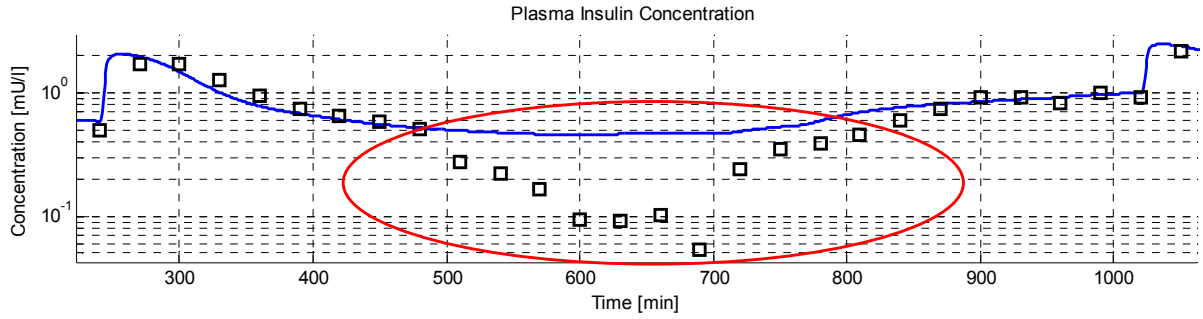

Figure 6: Simulated (straight blue line) and measured (boxes) plasma insulin concentration of Subject 11 (CSII protocol). The time-period of increased insulin clearance is marked with a red ellipse.

This behaviour could be caused by the binding characteristics of the insulin receptor. Receptor aggregation could decrease insulin binding affinity at higher insulin concentrations (Wanant and Quon 2000; Care and Soula 2011). Another reason could be the allosteric binding characteristics of the dimeric insulin receptor leading to negative cooperativity at high insulin concentrations (Kiselyov, Versteyhe et al. 2009). Both theories could support the insulin clearance phenomena observed in the clinical datasets.

In general, it has been acknowledged that in healthy subjects, and otherwise healthy subjects with diabetes, the insulin receptor is responsible for the largest fraction of insulin clearance (Duckworth, Bennett et al. 1998). Hepatic and peripheral insulin is both cleared during insulin signalling by insulin receptor binding and internalization. Even renal clearance is mediated by receptor based endocytosis of insulin (Duckworth, Bennett et al. 1998). Receptor binding characteristics are thus the main influence on the pharmacokinetics of insulin.

5.3.2 Dynamic changes in insulin clearance

The second observation, that the clearance of insulin is increased during a certain time period after a large insulin dose, is less well characterized and may be caused by intracellular signalling feedback mechanisms. The observations are depicted in Figure 7 and Figure 8. It is well known, that insulin stimulates a number of intracellular regulatory signalling cascades which, besides their direct glucoregulatory effect via GLUT4 translocation and hepatic glycogenesis, have various anabolic functions and may control the transcription of further regulatory proteins (Boron and Boulpaep 2008).

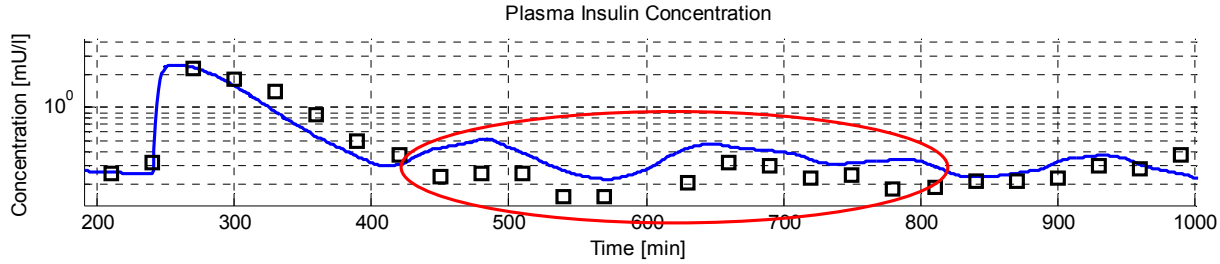

Figure 7: Simulated (straight blue line) and measured (boxes) plasma insulin concentration of Subject 1 (MPC protocol). The time-period of increased insulin clearance is marked with a red ellipse.

Similar regulatory effects could also influence the amount of insulin receptors in the tissue. As the insulin receptor is the main effector of insulin clearance, the regulation of receptor concentrations has a strong impact on insulin concentrations.

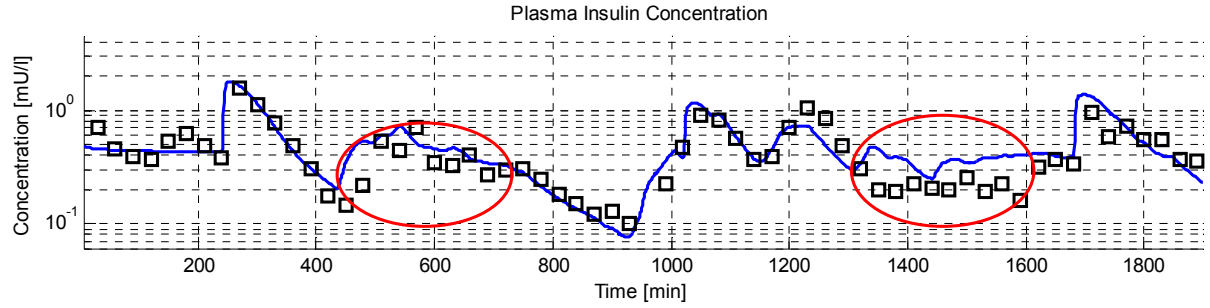

Figure 8: Simulated (straight blue line) and measured (boxes) plasma insulin concentration of Subject 2 (MPC protocol). The time-period of increased insulin clearance is marked with red ellipses.

The increased clearance of insulin after a given bolus could be caused by the up-regulation of overall receptor concentration or by the up-regulation of translocation of cytosolic receptors to the cellular membrane. Overall up-regulation of receptor concentration could be mediated by up-regulation of transcription. Up-regulation of receptor translocation could be either caused by a regulatory signal or by a mass-feedback akin to (Brannmark, Palmer et al. 2010) and the bolus shot causes the surface receptor concentrations to oscillate and thus overshoot following the bolus. The different theories on the relevant mechanisms are still being evaluated.

5.4 Limitations

However, it has to be taken into consideration that scientific knowledge on the detailed mechanisms of insulin clearance is still limited (Duckworth, Bennett et al. 1998) and that unknown effects could be the reason for these observations. As this behaviour could not be observed in every patient implies that it could be attributed to inter- or even intra-individual variability, the cause of which would also have to be determined.

Another possible explanation could be faulty insulin pump system (e.g. reduced pump precision at low infusion rates after giving boluses) or inter- and intra-individual variations in subcutaneous insulin absorption. Thus the regularity of occurrence of these observations has to be validated before a final conclusion can be drawn.

5.5 Outlook

Currently, no algorithms have been implemented to detect hypoglycaemia based on measurement data. However, the predictive power of the physiology-based models can be used to predict future trends of blood glucose in detail.

Also, using the physiology-based glucose-insulin model, a similar approach to evaluate pathophysiologic conditions of patients from the multi-parametric properties of insulin resistance could

be adapted. However, for this, extensive time-series datasets of blood glucose levels from patients with the conditions of interest would be necessary. To date, these datasets are not available.

The development of the physiology-based model of the glucose-insulin metabolism is documented in detail in deliverable D6-4-1. A validation of the model in a clinical context is planned. Beyond this effort, there are no efforts yet for additional model-based data analysis for risk assessment.

The observations presented here underline the complex mechanisms involved in the regulation of insulin concentrations and thus blood glucose levels. In any case, the complex mechanisms driving the PK of insulin are still not all well enough understood. Further research and a more detailed analysis of the relevant mechanistics, especially insulin clearance, will have to be conducted to explain the observed properties and ultimately improve the accuracy of insulin kinetics and thus the predictability of blood glucose levels.

6. Protocol for in-hospital insulin dosing prediction for T2DM at non-ICU

6.1 State-of-the-art and related work

Hyperglycemia in hospitalised patients with T2DM is a common and costly health care problem with profound medical consequences. Increasing evidence indicates that the development of hyperglycaemia during acute medical or surgical illness is not a physiologic or benign condition but is a marker of poor clinical outcome and mortality (Umpierrez et al 2002, Levetan & Magee 2000, Finney et al 2003, Umpierrez et al 2007). Observational studies in diabetic subjects admitted to general medical and surgical areas have shown that poor glycaemic control is associated with prolonged hospital stay, infection, disability after hospital discharge and death (Umpierrez et al 2002, Clement et al 2004, Pomposelli et al 1998).

In the critical care setting, a variety of continuous insulin infusion protocols have been shown to be effective in achieving glycaemic control, with low rate of hypoglycaemic events (Inzucchi 2006, Goldberg et al 2004). Prospective randomized trials in postsurgical patients have shown that improved glycaemic control reduces short- and long-term mortality, rates of multi-organ failure and systemic infections, and length of hospitalization (Van den Berghe et al 2001 and 2007). However, in a study with mixed surgical-medical intensive care patients results could not be confirmed (NICE Sugar Study Group 2009).

In patients with diabetes admitted to general medicine wards hyperglycaemia is commonly not well addressed (Umpierrez et al 2002, Levetan & Magee 2000, Schnipper et al 2006, Umpierrez & Maynard 2006). Although several guidelines for treatment regimen for outpatient management of T2DM have been defined (Sakharova & Inzucchi 2005, Inzucchi 2006), no clear definitions of treatment regimen have been found for the establishment of glycaemic control of hospitalised patients (Clement et al 2004, ADA 2010).

The consensus panel of the American Diabetes Association reviewed research together with the original investigators to formulate standards for diabetes management in the hospital. The panel concluded that hospitalised patients should have a target glycaemic premeal / fasting level of <140 mg/dL (7.8 mM) and that insulin, whether administered intravenously or subcutaneously is the primary means of effective glycaemic control in the hospital setting (ADA 2010).

Reports from academic institutions have shown that most patients are treated with sliding scale regular insulin (SSI) and that basal insulin is prescribed in less than half of the patients (Schnipper et al 2006). A recent audit of glycaemic control in two general wards of the University Hospital of Graz has shown that hospitalised patients with T2DM have a mean blood glucose level above the recommended target range. The analysis of 50 consecutive patients, who were treated with insulin at the general wards of endocrinology and cardiology, revealed a mean blood glucose level of 167 mg/dl. No difference between admission and discharge blood glucose values was observed, indicating an insufficient insulin titration process throughout the hospital stay[4].

In contrast, use of a basal bolus insulin treatment protocol in a multi-centre randomized trial has shown that in general medicine patients with non-insulin treated T2DM diabetes glycaemic control was improved without increasing the risk of severe hypoglycaemia compared with SSI regimen (Umpierrez et al 2007). In an another study, T2DM patients under previous treatment with diet, oral agents and/or insulin the same basal bolus insulin titration protocol was equivalent with NPH and regular insulin twice daily (Umpierrez et al 2009). Most recently, this protocol demonstrated in a randomized study design that it could not only improve glycaemia, but also reduced the incidence of a composite of postoperative complications (wound infections, pneumonia, bacteraemia, respiratory and acute renal failure) (Umpierrez et al 2011). Thus, this protocol represents the currently published best practice (PBP) to manage glycaemia in the non-intensive care unit patients with T2DM.

One aim of within REACTION is to develop an insulin titration protocol based on PBP to provide decision support for optimal glucose management for nurses and physicians at a general ward. This

[4] One possible explanation for these results is that many hospitals are having difficulty in maintaining the necessary number and quality of nurses to monitor general ward diabetic patients, understand the observations and arrange appropriate interventions despite improved monitoring tools.

protocol serves as basis for a mobile workflow and insulin dosing support software application in hospitals.

6.2 Methods

The development of an insulin dosing protocol for a general ward started with a workflow analysis of the current treatment situation of patients with T2DM (mainly) at the ward of Endocrinology and the ward of Cardiology at Graz. In the first step the treatment protocol of 50 patients (age 71.5±12.3 years, BMI 28.3±6.1kg/m2, 22 men/28 women, 3x T1DM, 47x T2DM, HbA1c 7.8±1.5%) has been retrospectively analysed regarding characteristics of the glucose management process: number BG measurements, mean BG, BG measurement in the ranges 70-140mg/dl, 70-180mg/dl, <60mg/dl and >300mg/dl. The data were analysed per population, patient-day and patient-stay. In addition, glucose and insulin were analysed for the admission and discharge period (Neubauer et al, 2011). Figure 9 shows the current status of BG management. It shows that blood glucose values above target level (140 mg/dl) and no reduction of the BG level during the hospital stay can be observed. These outcomes at Graz which are comparable to other hospitals (Umpierrez et al 2002, Levetan & Magee 2000, Finney et al 2003, Umpierrez et al 2007) is not satisfactory and acted as an indicator that improvements of workflow and therapy were needed.

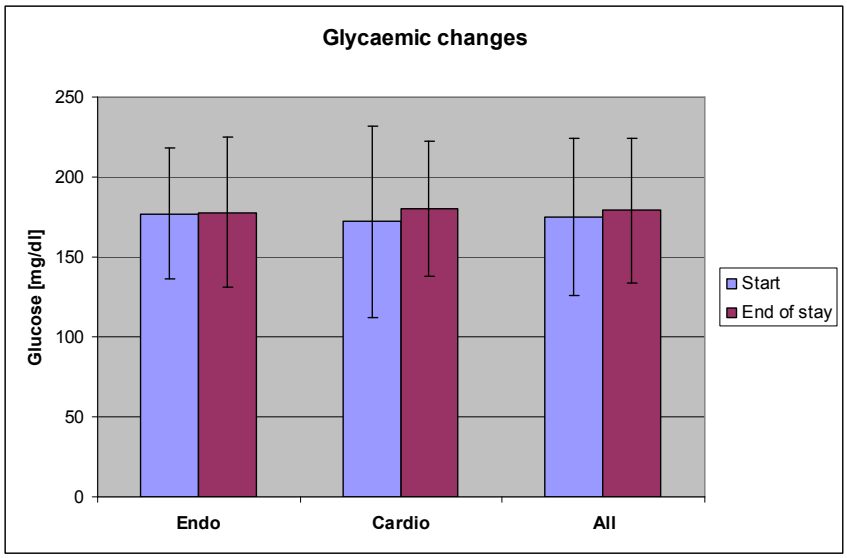

Figure 9: Blood Glucose levels in the beginning and at the end of a hospital stay

In the second step a workflow analysis for the treatment of patients with T2DM has been conducted. Therefore, nurses and physicians have been questioned concerning the treatment process, and available treatment schemes have been analysed. The main outcome of this survey was:

- No SOPs for glycaemic control are defined
- No standardised workflow of glycaemic control
- Different physicians use different treatment/therapy regimes
- SOPs are needed (definition of target range, insulin dosing, correctional scheme)
- Advice by the physicians should be available
- Immediate action after blood glucose measurement
- Training of physicians and nurses

Based on the results of the current status analysis, a target analysis has been performed. A simplified process diagram has been composed and served as starting point for the insulin dosing protocol for REACTION (see Figure 10).

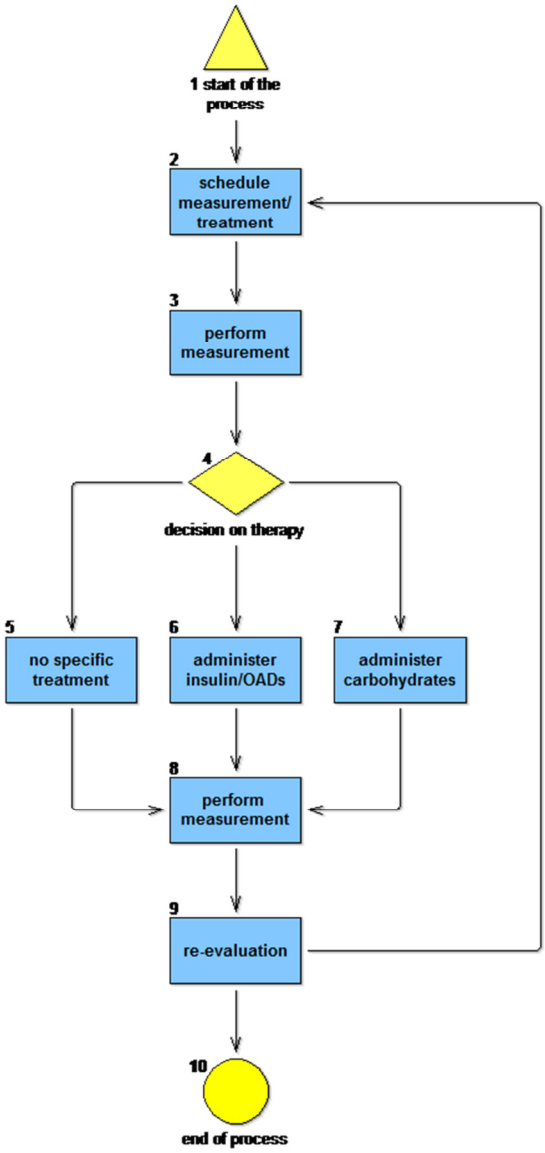

Figure 10: Workflow description of glycaemic management of a general ward

An in-deep literature research for insulin dosing protocol for patients with type 2 diabetes has been performed (Sakharova & Inzucchi 2005, Inzucchi 2006, Umpierrez et al 2007, Umpierrez et al 2009, Umpierrez et al 2011) to overcome the problems stated above. Following general requirements to the insulin dosing protocol have been identified:

- safe and effective
- simple – easy to handle and traceable for clinical personnel
- fit into workflow for patient treatment at a general ward
- possibility to integrate into mobile clinical workflow system

As starting point the RABBIT-2 trial (Umpierrez et al 2009, Umpierrez et al 2011) has been chosen. The protocol has been adapted in an iterative discussion and prototyping process by the interdisciplinary team of medical diabetes experts, nurses and technicians of MUG and MSG. All decisions are affected by the requirements of putting the protocol into an electronic mobile system for glucose management and decision support.

Various adaptions and extensions of the protocol have been done:

- Extension of unknown/incomplete actions (e.g. parameters for daily dose calculation)

- Consideration of possible side effects (e.g. missing blood glucose values)
- Transformation of protocol to a deterministic, robust algorithm for the electronic system
- Integration into workflow and insulin dosing support system for a general ward

6.3 New knowledge gained from literature - Results

Figure 12 presents the developed workflow for the treatment of patients with T2DM at a general ward. It incorporates workflow requirements of the clinical personnel as well as a decision support for insulin dosing. The decision support helps nurses and clinicians to find the proper dosage for their patients.

The process starts with the admission of a patient at the general ward. Relevant patient and treatment parameters are automatically transferred from the hospital information system to the developed software solution. Only patients who fulfil pre-defined medical parameters (e.g. diagnosis: T2DM, creatinine level below a specific value, no gestational diabetes) will be enrolled for the glucose management and decision support system. During the enrolment process various initial parameters related to the medical status and the therapy of a patient will be specified by the clinician.

The decision support provides at this state of the workflow the clinician with an initial daily insulin dose based on age, weight and creatinine level of the patient.

After therapy initialization four times per day blood glucose is measured before meal and the decision support suggests the proper bolus units based on former blood glucose measurements. If the dosage is acceptable for the decision maker the insulin units are administered.

The total daily insulin dosage is composed by the basic bolus insulin and the basal insulin. The basal insulin is admitted once a day at midday. Depending on the current blood glucose value supplement bolus insulin is calculated by the decision support. Figure 11 shows the composing of the different insulin (basal, bolus, supplement) provided as dosage advice by the decision support.

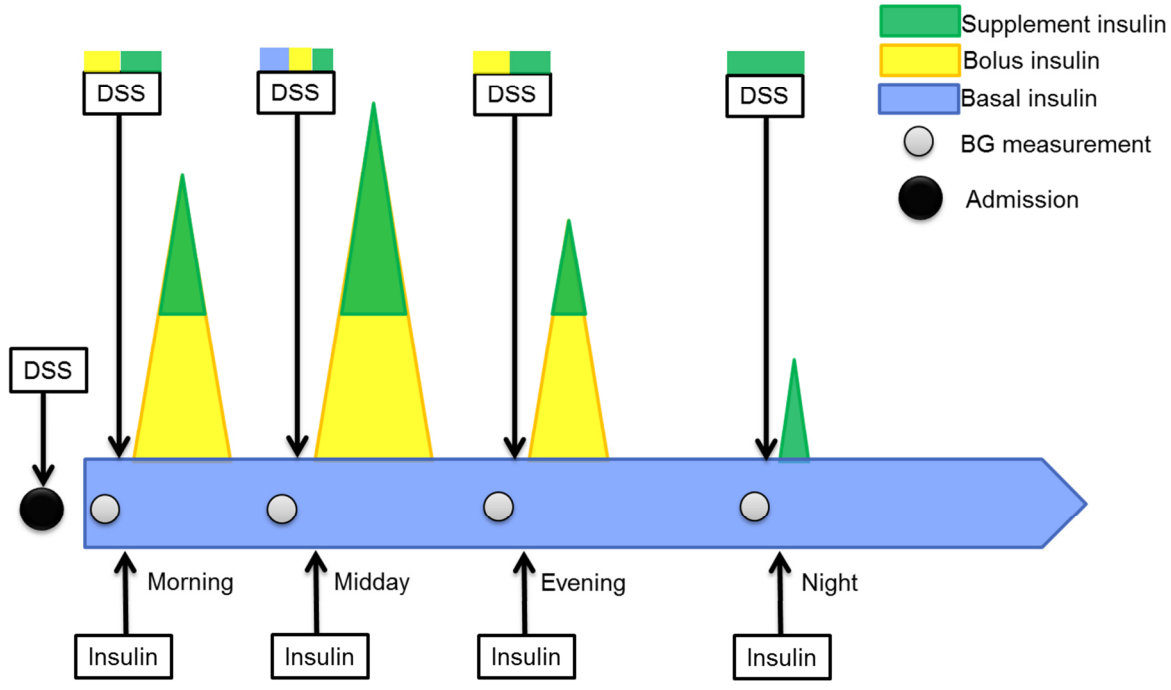

Figure 11: Decision support for insulin dosing

Once per day, at the ward round, an adjustment of the therapy is performed and a new daily insulin dose will be suggested by the decision support. The new dosage is based on the old dosage and former blood glucose values (morning and evening values of the previous day) of the patient. This daily insulin dosage sets the new basal and bolus insulin dose.

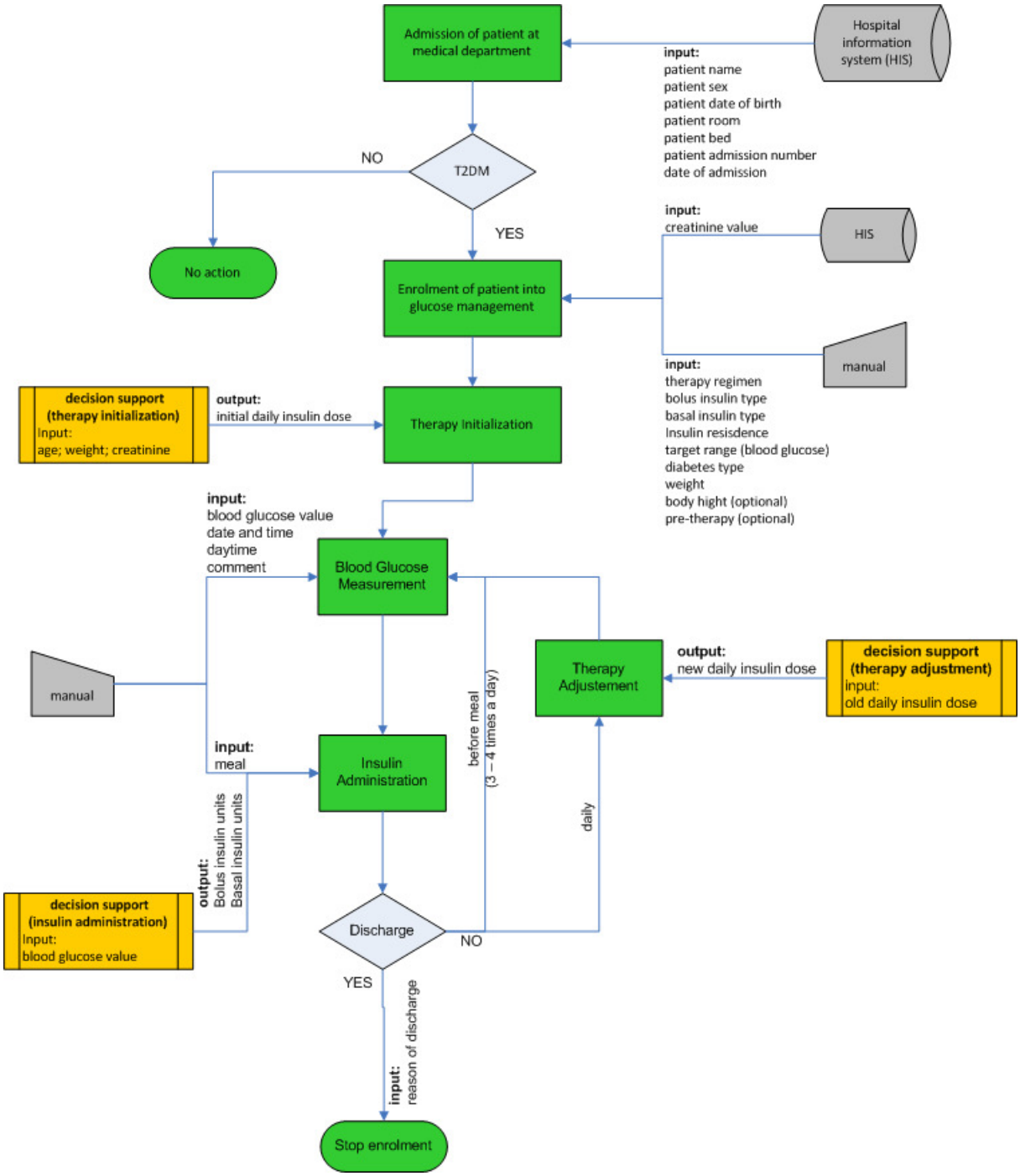

Figure 12: Workflow diagram for glucose management based on Basal/Bolus regimen for patients with T2DM

Based on the developed workflow and insulin dosing protocol a tablet-based mobile client/server software application has been developed. The software supports clinical personal directly at the point of care to improve the treatment of non-ICU patients with T2DM. The following screenshots show the main functions of the software and reproduce the process diagram in Figure 12 electronically.

Figure 13 presents the input parameters for the initial insulin dose calculation after a patient has been admitted at the ward. Based on the weight and the creatinine value of the patient a daily insulin dose suggestion is provided by the dosing protocol. Figure 14 presents the result of the calculation. The system provides the daily insulin dose separated into basal and bolus insulin for the current day. Additionally, the bolus insulin is divided into the doses for the times of the day. A responsible physician

can now prescribe the suggestion insulin dose or (s)he can change the dose if there are doubts on the suggestion. The type of insulin and other important parameters for the treatment have been entered during the enrolment process (for details please see *D2.6 Prototype Application Specification Appendix*)

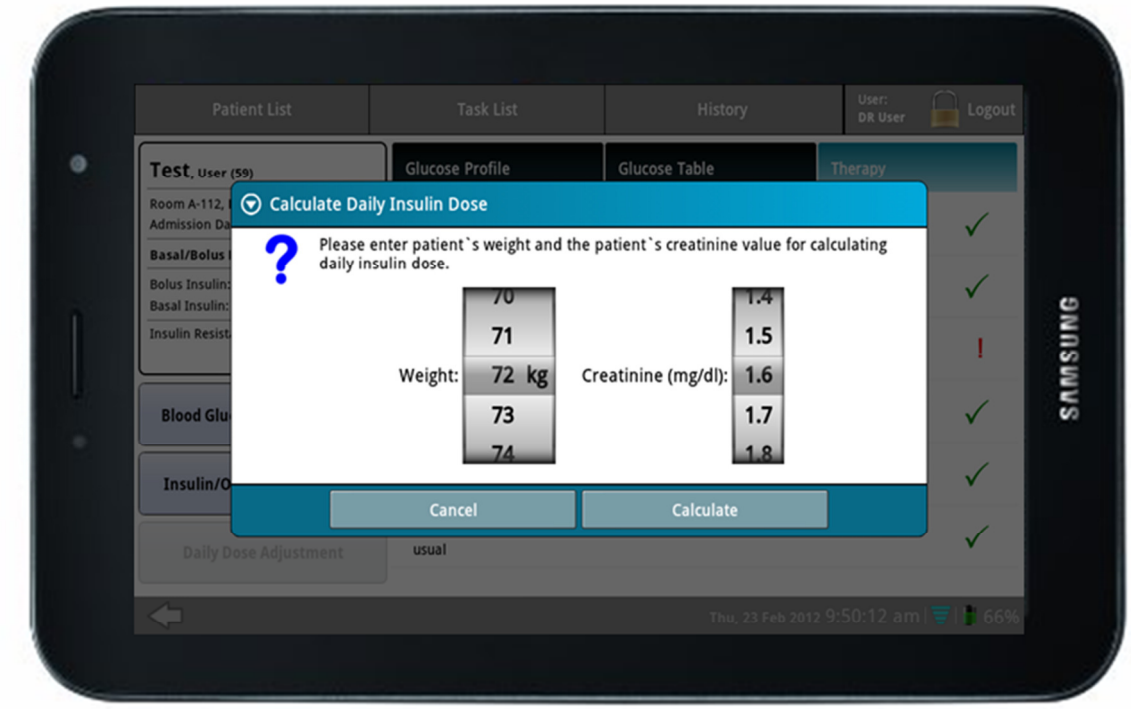

Figure 13: input parameters for initial calculation of daily insulin dose

Figure 14: output of initial daily insulin dose calculation

Figure 15 shows the insulin dose suggestion for the morning. Bolus insulin is administered before each meal. In the presented screenshot only bolus insulin composed of 6 basis bolus units plus 4 units

of supplement bolus are suggested and have to be checked and finally administered. Basal insulin is administered after the ward round at midday (where the therapy adjustment has to be done).

Figure 15: pre-meal insulin dose suggestion (morning) by the decision support

During the ward round, therapy adjustment has to be performed by the clinicians each day. The result of the decision support is the suggestion of a new daily insulin dose which has to be approved by a responsible clinician. Figure 16 shows the result of the therapy adjustment for the test user.

Figure 16: Therapy adjustment – new daily insulin dose

Finally the results of the blood glucose measurements and the insulin administration are displayed on the main screen of the glucose management system – see Figure 17.

Figure 17: main functions and data visualisation

6.4 Limitations

Insulin dosage suggestion is mainly based on blood glucose values and does consider the complex physiological situation of a specific patient in a basic way (e.g. three categories for estimation of insulin sensitivity). Consequently, changes in the physiology of the patient or accompanying medical events (e.g. fever, vomiting, and other reasons for insulin resistance) are not modelled in the protocol. The physiology-based model of Bayer attempts to give answers to these questions.

The current version of the protocol is restricted to a quite fixed workflow of spot measurements on defined times of the day. Efforts go to separate the protocol from the workflow system (configurable workflow system) to ease the exchange of the underlying protocol.

6.5 Outlook

Currently the insulin dosing protocol is tested on paper in a clinical study. Intermediate results are promising and it can be expected that the protocol will perform well. The clinical study will be finished approx. in spring 2012. In the next step a clinical study with the software solution will be performed in summer 2012. Based on the already existing results of two usability tests the usability, safety and performance of the software solution will be tested at the ward of Endocrinology at the Medical University of Graz. This study provides first insights in the prospects of success of the in-hospital workflow and insulin dosing support system.

Furthermore, the incorporation of the physiology-based model of Bayer into the in-hospital workflow and insulin dosing support system is being considered. The exact modelling of patient physiology and the possibilities for hypo-detection may improve the quality of the system.

7. Text mining for decision support in diabetes treatment

7.1 State-of-the-art and related work

Besides the structured databases like the EPR or monitoring databases, a lot of natural text documents exist. In the medical domain, these include for example literature (scientific articles for professionals and educational texts for patients), textual parts of electronic health records and medical documents like discharge letters. Information retrieval or knowledge discovery from these sources require different methods which are often referred as text mining. Text mining has been defined as the non-trivial discovery process for uncovering novel patterns in unstructured text. Text mining techniques have been advanced by using techniques and methods from information retrieval, natural language processing, data mining, machine learning, and statistics. However, some old methods are rarely used nowadays. For example, while Bayesian models and hierarchical clustering were widely used in the early days, more advanced machine learning methods, such as artificial neural networks, support vector machines, and semantic-based clustering algorithms, have been applied in recent years. The next section provides an overview of typical text mining applications based on the reviews of Spasic et al. (2005), Cohen and Hersh (2005) and Yoo and Song (2008).

A common text mining method is document clustering. Basically, document clustering is to group unlabelled documents into meaningful document clusters whose documents are similar to one another within the cluster, without any prior information about the document set. In order to measure similarities between documents, they are often represented based on the vector space model. In this model, each document is represented as a high dimensional vector of words'/terms' frequencies. A number of document clustering approaches has been developed over several decades. Most of these document clustering approaches are based on the vector space representation and apply various clustering algorithms to the representation. Many approaches can be categorized as hierarchical or partitional clustering. There are attempts to apply semantics to the clustering algorithm by mapping terms in documents to ontology concepts and then aggregating concepts based on the concept hierarchy.

Text classification is a technique to automatically determine the category that a document or part of a document belongs to, based on the particular topics or characteristics of interest that a document contains. Accurate text classification systems can be especially valuable for example to database curators, who may have to review many documents to find a few that contain the kind of information they are collecting in their database. Classification methods primarily rely on supervised machine learning techniques. Features used for classification are not specified explicitly by the user; instead the user only provides a set of documents that contain the characteristics of interest, which is known as the positive training set, and another set that does not contain the characteristics, which is known as the negative training set.

The goal of a text classification system is to assign the class labels to the new unseen documents. Classification takes place in two different phases, namely model construction, or also widely known as the training phase, and model usage, also known as the prediction phase. In the training phase, given a set of positively labelled sample documents, the goal is to automatically extract features relevant to a given class. Hence, it would help distinguish the positive documents from the negative ones. Once such features have been identified then those features should be applied to the candidate documents using some kind of decision-making process. In the model usage phase, the accuracy of the model is checked and then used to classify new unseen data.

The huge volume of the biomedical literature provides a promising opportunity to induce novel knowledge by finding novel connections among logically-related medical concepts. For example, Swanson introduced Undiscovered Public Knowledge (UDPK) model to generate biomedical hypotheses from biomedical literature such as MEDLINE [Swanson 1986]. According to Swanson, UDPK is "a knowledge which can be public, yet undiscovered, if independently created fragments are logically related but never retrieved, brought together, and interpreted". The UDPK model formalizes a procedure to discover novel knowledge from biomedical literature as follows: Consider two separate sets of biomedical literature, BC and AB, where the BC document set discusses biomedical concepts B and C and the AB document set discusses biomedical concepts B and A. However, none of the documents in the BC or AB sets primarily discusses biomedical concepts C and A together. The goal of the UDPK model is to discover some novel connections between a starting concept C (e.g., a disease) and target concepts A (e.g., possible medicine or treatments to the disease) by identifying

biomedical concept B (called a bridge concept). For example, Swanson discovered that fish oils (as concept A) could be a potential medicine for Raynaud disease (as concept C) by identifying the bridge concept (as concept B) blood viscosity. Swanson's UDPK model can be described as a process to induce "C implies A", which is derived from both "C implies B" and "B implies A"; the derived knowledge or relationship "C implies A" is not conclusive but, rather, hypothetical. The goal is to uncover previously unrecognised relationships worthy of further investigation.

Another type of text mining is information extraction. In terms of what is to be extracted by the systems, most studies can be broken into the following three major areas: 1) named entity extraction (NER), named entities may include proteins or genes, 2) relation extraction whose main task is to extract relationships among entities, and 3) abbreviation extraction. Most of these studies adopt information extraction techniques, using a curated lexicon or natural language processing for identifying relevant tokens such as words or phrases in text.

Named entity recognition is a crucial step in extracting more complex types of information (i.e. facts and events). The idea is that recognising biological entities in text allows for further extraction of relationships and other information by identifying the key concepts of interest and allowing those concepts to be represented in some consistent, normalised form. This task has been challenging for several reasons. First, a complete dictionary for most types of biological named entities does not exist, so simple text matching algorithms do not suffice. In addition, the same word or phrase can refer to a different thing depending on the context. Conversely, many biological entities have several names. Biological entities may also have multi-word names, so the problem is further complicated by the need to determine name boundaries and resolve overlap of candidate names. With the large amount of genomic information being generated by biomedical researchers, it should not be surprising that in the genomics era, much of the work in biomedical NER has focused on recognising gene and protein names in free text. The approaches generally fall into three categories: lexicon-based, rules-based and statistically based.

The goal of relationship extraction is to detect occurrences of a pre-specified type of relationship between a pair of entities of given types. While the type of the entities is usually very specific (e.g. genes, proteins or drugs), the type of relationship may be very general (e.g. any biochemical association) or very specific (e.g. a regulatory relationship). Several approaches to extracting relations of interest have been reported in the literature and are applicable to this work. Manually generated template-based methods use patterns (usually in the form of regular expressions) generated by domain experts to extract concepts connected by a specific relation from text. Automatic template methods create similar templates automatically by generalising patterns from text surrounding concept pairs known to have the relationship of interest. Statistical methods identify relationships by looking for concepts that are found with each other more often than would be predicted by chance. Finally, NLP-based methods perform a substantial amount of sentence parsing to decompose the text into a structure from which relationships can be readily extracted. In the current genomic era, most investigation of this type has centred around relationships between genes and proteins. It is thought that grouping genes by functional relationships could aid gene expression analysis and database annotation. Several researchers have investigated the extraction of general relationships between genes.

Paralleling the growth of the increase in biomedical literature is the growth in biomedical terminology. Because many biomedical entities have multiple names and abbreviations, it would be advantageous to have an automated means to collect these synonyms and abbreviations to aid users doing literature searches. Furthermore, other text-mining tasks could be done more efficiently if all of the synonyms and abbreviations for an entity could be mapped to a single term representing the concept. Most of the work in this type of extraction has focused on uncovering gene name synonyms and biomedical term abbreviations. Van der Zanden (2010) applied information extraction methods to textual electronic patient records in order to evaluate their quality.

Information retrieval is extensively used by biomedical experts to locate relevant information (most often in the form of relevant publications) on the Internet. Apart from general-purpose search engines, many IR tools have been designed specifically to query the databases of biomedical publications such as PubMed. It is particularly important in biomedicine not to restrict IR to exact matching of query terms, because term ambiguity and variation phenomena may cause irrelevant information to be retrieved (low precision) and relevant information to be overlooked (low recall). Adding semantics would be particularly useful, e.g. the hierarchical organisation of ontologies and relations between the described concepts (and through them the corresponding terms) can be used to constrain or relax a search query and to navigate the user through huge volumes of published information.

Many current semantic information retrieval solutions use ontologies [see e.g. Vallet et al. (2005) and Styltsvig (2006)], but the construction of a comprehensive ontology even for a narrow domain requires much effort. We have analyzed the public ontology repositories (like the Open Biomedical Ontologies or the BioPortal) and the scientific literature and concluded that currently no comprehensive ontology exists that satisfactorily covers all diabetes-related areas relevant for information retrieval. However, there are some ontologies that cover parts of the focus domain. Moreover, some publications (Shahar 1996, Villarreal 2009) use ontologies to solve diabetes-related problems, and publish parts of these ontologies in different forms.

Among existing biomedical domain ontologies, the closest to satisfying some of the requirements is the Chronic Disease Ontology (Verma 2009). Unfortunately, in its present state it can serve at best as a starting point for the development of the biomedical part of a diabetes domain ontology. First of all, it is rather incomplete: e.g. the ontology does not contain the complications of diabetes like neuropathy, and none of the diabetes signs and symptoms that are present in the ontology are connected to the instance “Type-2 diabetes mellitus” with the “associated_with” relation, the intended role of which is exactly to link chronic diseases with their signs and symptoms. Secondly, and more importantly, a number of conceptual/modelling problems can be raised regarding its treatment of the domain. One of the gravest conceptual problems is the frequent misuse of the is_a metarelation: the ontology often connects classes via this relation that obviously cannot have the subclass/superclass relation to each other, e.g. it states that the “Blood test” class is a subclass of the “Disease” class and the “Patient_group” class is the subclass of the “Human” class.

7.2 Aim

If we would plan to develop an ontology-based semantic search solution, the existing biomedical ontologies partly covering the diabetes domain could be used only with extensive conceptual re-engineering and with the addition of large amounts of missing information. This would require much efforts and close cooperation between ontology experts and health professionals. The necessary resources are not provided by the REACTION project.

As an alternative, ALL has developed a semantic information retrieval technology which relies on the users' relevant background knowledge in the form of user-specified “ontology capsules”, and therefore does not require pre-built, domain-specific semantic resources (i.e. ontologies) in order to achieve high precision values. With this approach queries can be specified via natural language expressions and sentences, and the system searches for text fragments in the document set that have approximately the same meaning (semantic content) as the query.

This technology opens up the possibility of developing efficient and comfortable search systems for repositories containing medical documents in natural languages, e.g. scientific papers or natural language text fragments entered in electronic patient records. It also enables search in medical data repositories containing large amounts of natural language documents, e.g. a repositories of hospital discharge letters.

The technology is capable of supporting multilingual search scenarios, since queries are transformed into language-independent semantic representations that can be matched with texts in different languages. In addition to IR, the approach can also be used for information extraction (IE), e.g. a search for medicines usable for the treatment of a certain set of symptoms will return the list of concrete medicines (that is, the list of the expressions by which they are referred to) found in the relevant documents.

Our aim is to implement this technology to be used in the diabetes domain within the REACTION platform. The most probable users are physicians, but other health professionals and eventually also patients may be considered. The developed component will support its users in finding the most appropriate information which is necessary to make decisions. The solution can be used at every level of health care, including primary care as well as hospitals, thus it might be useful for both clinical partners CHC and MUG.

There are various databases which are potentially useful for information retrieval and/or extraction:

- The Electronic Patient Record, containing structured and non-structured (textual) information about the patient’s health status and history

- Guidelines (possibly including patient guidelines) mostly available as natural language documents
- Evidence-based medicine repositories (e.g. the Cochrane library) containing high-quality scientific articles

The exact goals can be determined only after discussion with the end-users.

7.3 Method

This section describes the methodologies used for the implementation of the semantic information retrieval (SIR) component. Figure 18 provides an overview on the components of the SIR.

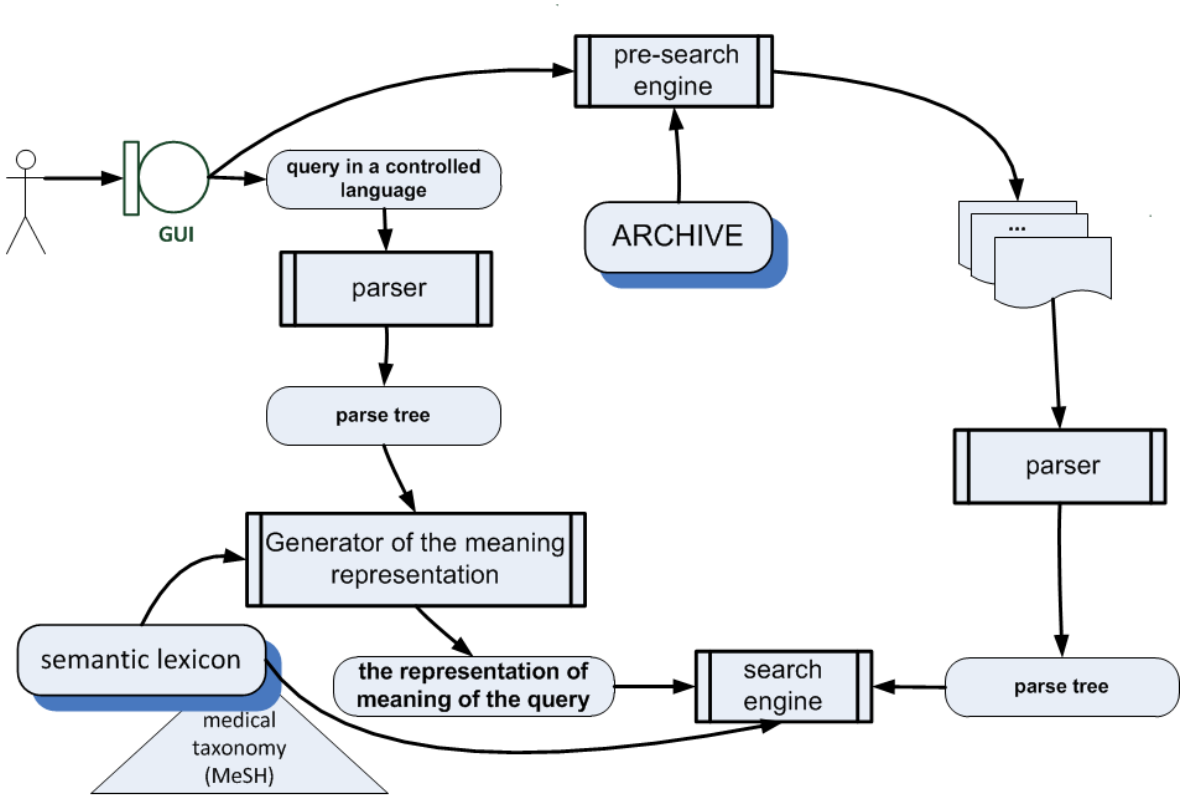

Figure 18: Structure of SIR

The query is a structure of noun phrases and sentences expressed in a controlled natural language which ensures the disambiguity of the query. The sentences can be connected by boolean operators.

The Parser creates a graph (DAG – directed acyclic graph) representation of the syntactic structure of the query and of segments of natural language texts. It consists of several modules as it is usual in natural language processing (NLP). Some of them are standard NLP programs; however, the most important components are made for the semantic search according to special requirements. Parsing the query or segments of documents is somewhat different, because the inputs are different. In the first case a pre-processed and structured version is the input, in the second case the input is the text itself. Therefore while in the first case the correctness of parsing is ensured, in the second case it is not. However, problems in the parsing of text segments do not cause failures in the search, because of the strategy built into the search engine.

The Semantic Lexicon stores the lexical units (words) with their semantic properties. It consists of several components. There are separate but connected storage for the predicative words (words expressing events, relations, mainly verbs) and for non-predicative ones. The reason is that different resources store them. The predicative words are obtained from FrameNet; however, the FrameNet corpus has to be completed substantially. WordNet is used for the non-predicative words. Above these

stores of words there are ontology capsules acting as upper ontology segments. The medical terminology is obtained from MeSH. If necessary for the intended usage, it can be extended with other diabetes-specific terms not included in the terminology. The words are grouped into synonym-sets (synsets). We use the notion of being synonym in a less strict way than it is used in linguistics: two words are synonym, if they refer to similar situation or item. This definition is adequate for information retrieval. With the predicative words their valence patterns are also represented, including semantic features (role relations) of the valence units. The role relations are connected to the relations defined in the ontology capsules.

The Generator of the Meaning Representation is responsible for building a meaning representation of the query from the result of the Parser. The meaning representation is a DAG similar to the syntactic structure; however the nodes are labelled by the synsets of the actual words, the edges are labelled by the semantic relations between the words. The synsets and the semantic relations come from the Semantic Lexicon. The words determine the possible synsets, different senses of a word belong to different synsets, they are called lexical units. The syntactic relations between the words in the sentences are matched to the valence patterns of the lexical units, and the semantic relations are rendered to valence units (the elements of the valence patterns). This matching and the application of semantic restrictions help to select the most probable LUs for the words.

The Pre-search Engine substantially reduces the amount of documents to be processed by a preliminary key-word based search. If data base of EPRs are searched, it executes the data base search based on numerical and coded data.

Finally, the Search Engine finds the phrases that may have similar meaning as the query (or its part) has. As the above figure shows, the texts are syntactically parsed, but no meaning representation is generated. In the search we only test whether the syntactic structure may fit into the meaning representation of the query.

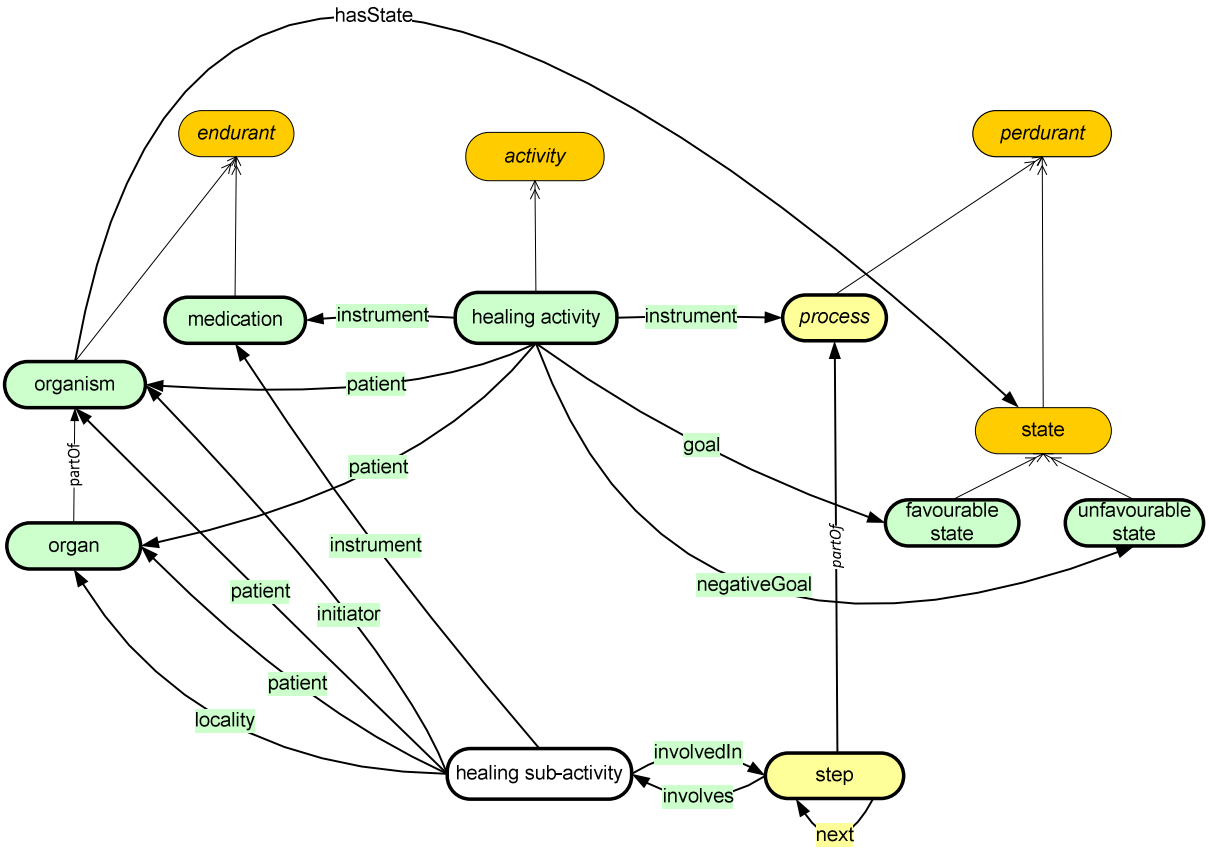

Figure 19: Example of an ontology capsule

This methodology was developed taking into account the recent scientific literature. Relevant publications for readers interested in more details include: Baker et al. (1998); Baumgartner and

Burchardt (2004); Burchardt and Frank (2005); Giuglea and Moschitti (2006) and Scheffczyk et al. (2006).

7.4 Working plan

ALL has already designed the technology for semantic information retrieval, and the implementation of general-purpose components (see also Figure 18) has been started. The requirements for the use of this technology within the REACTION project need to be determined, and after that, the specific components need to be implemented or adapted.

The implementation of the Semantic Information Retrieval component for diabetes can be divided into the following eight tasks:

1. Determination of the user requirements
2. Tuning the controlled query language to the requirements
3. Tuning the Parser and the Meaning Representation Generator to the requirements
4. Implementation of the Search Engines
5. Preparation of the necessary semantic resources: Semantic Lexicon, domain-specific situations and frames
6. Integration of the components
7. Testing by end-users
8. Evaluation of the test results and modification of the system

It is important to note that the tasks 1 and 7 require the active contribution of end users, i.e. health professionals. Task 1 includes the following parts:

- Determination of expected users (e.g. physicians, nurses, patients)
- Determination of databases in which the queries need to be performed (e.g. EPR, guidelines, Cochrane database)
- Extraction of typical queries to be answered by the component. A promising opportunity to gather such knowledge is for example to study the case conferences planned by CHC.

This is a fundamental task because the component can be tailored to the diabetes domain and the user needs only if this task is properly performed. For example, the necessary semantic resources depend on the intended use of the component. Feedback from users in task 7 is also indispensable in order to fine-tune the developed component.

The timing of these tasks is illustrated in Figure 20. The prototype will be prepared by the end of year 3, after the successful completion of tasks 1 to 6. The component will be finalized after testing and implementation of the modifications suggested by the test results.

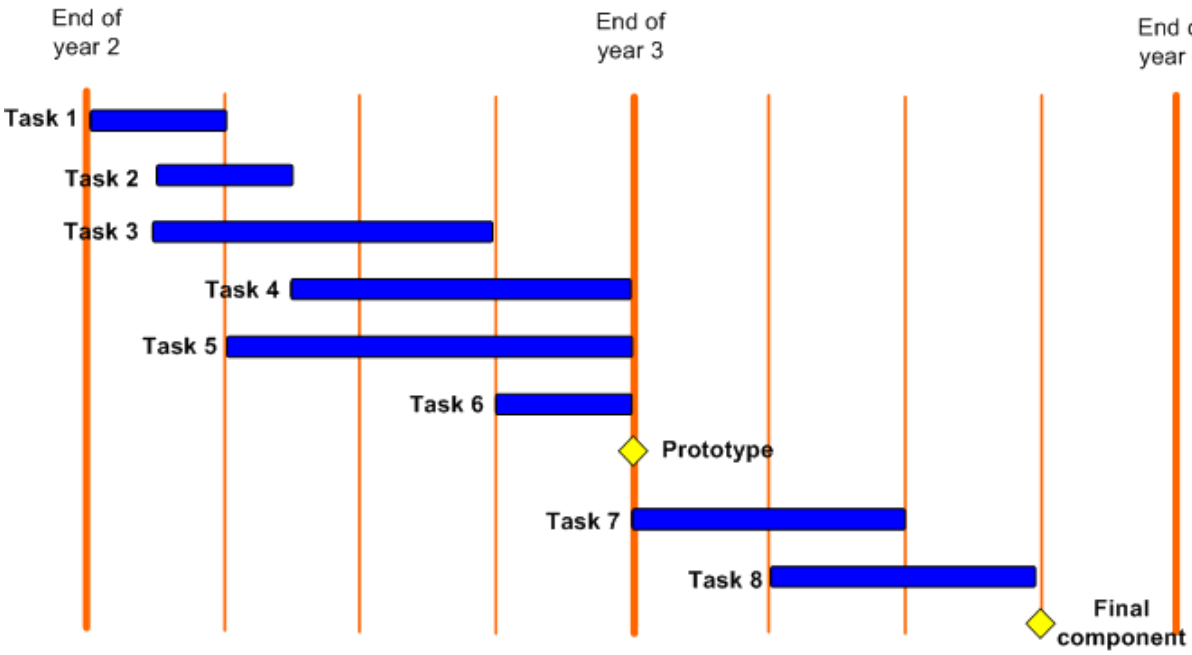

Figure 20: Timing of the implementation tasks

7.5 Expected results

Information retrieval is a useful tool for locating relevant information in repositories of textual documents, e.g. guidelines or evidence-based medical articles. In contrast to general-purpose, keyword based search engines, adding semantics helps to retrieve more precise results. The aim of our work is to implement the semantic informational retrieval method developed by ALL for the diabetes domain. This method has the advantage that it does not require a large pre-built ontology, instead, it uses the relevant background knowledge of the users in the form of “ontology capsules”.

The expected result of the development is to implement a semantic information retrieval component which is tailored to the requirements of the diabetes domain. More precisely, it will be able to retrieve and/or extract information from one or more of the following sources: the electronic patient record, guidelines and evidence-based literature (e.g. the Cochrane library). The first step of development must be the determination of the exact requirements with the inclusion of end users, e.g. to decide in which repositories should the search be performed. These decisions will influence the implementation details such as the necessary semantic resources. The prototype of the component is expected to be prepared by the end of year 3. This will be followed by a testing period. The prototype will be refined based on the test result.

8. References

[Aliferis et al. (2009)] C. F. Aliferis, A. Statnikov, I. Tsamardinos, J. S. Schildcrout, B. E. Shepherd and V. B. Bajic. Factors Influencing the Statistical Power of Complex Data Analysis Protocols for Molecular Signature Development from Microarray Data. PLoS ONE, 4(3):e4922, 2009

[Aliferis et al. (2010a)] C. F. Aliferis, A. Statnikov, I. Tsamardinos, S. Mani and X. D. Koutsoukos. Local Causal and Markov Blanket Induction for Causal Discovery and Feature Selection for Classification Part I: Algorithms and Empirical Evaluation. Journal of Machine Learning Research, Special Topic on Causality, 11:171-234, 2010

[Aliferis et al. (2010b)] C. F. Aliferis, A. Statnikov, I. Tsamardinos, S. Mani and X. D. Koutsoukos. Local Causal and Markov Blanket Induction for Causal Discovery and Feature Selection for Classification. Part II: Analysis and Extensions. Journal of Machine Learning Research, Special Topic on Causality, 11:235-284, 2010

[Cederholm et al. (2008)] Jan Cederholm, Katarina Eeg-Olofsson, Björn Eliasson, Björn Zethelius, Peter M Nilsson, Soffia Gudbjörnsdottir, and Swedish National Diabetes Register. Risk prediction of cardiovascular disease in type 2 diabetes: a risk equation from the Swedish National Diabetes Register. *Diabetes Care*, 31 (10): 2038–2043, Oct 2008.

[Clarke et al. (2004)] P. M. Clarke, A. M. Gray, A. Briggs, A. J. Farmer, P. Fenn, R. J. Stevens, D. R. Matthews, I. M. Stratton, R. R. Holman, and U. K. Prospective Diabetes Study (UKDPS) Group. A model to estimate the lifetime health outcomes of patients with type 2 diabetes: the United Kingdom Prospective Diabetes Study (UKPDS) Outcomes Model (UKPDS no. 68). *Diabetologia*, 47 (10): 1747–1759, Oct 2004.

[DCCT Group (1993)] The Diabetes Control and Complications Trial Research Group. The Effect of Intensive Treatment of Diabetes on the Development and Progression of Long-Term Complications in Insulin-Dependent Diabetes Mellitus, *New England Journal of Medicine*, 329:977-986, 1993

[DCCT Group (1995)] The Diabetes Control and Complications Trial Research Group. Adverse Events and Their Association With Treatment Regimens in the Diabetes Control and Complications Trial. *Diabetes Care*, 19:361-376, 1995

[DCCT Group (1996)] The Diabetes Control and Complications Trial Research Group. Effects of intensive diabetes therapy on neuropsychological function in adults in the Diabetes Control and Complications Trial. *Annals of Internal Medicine*, 124(4):379-88, 1996.

[EDIC Group (1999)] Epidemiology of Diabetes Interventions and Complications Study Group. Epidemiology of Diabetes Interventions and Complications (EDIC). Design, implementation, and preliminary results of a long-term follow-up of the Diabetes Control and Complications Trial cohort. *Diabetes Care*, 22 (1): 99–111, Jan 1999.

[ETDRS Group (1991)] Early Treatment Diabetic Retinopathy Study Research Group. Fundus photographic risk factors for progression of diabetic retinopathy. ETDRS Report Number 12. *Ophthalmology*, 98:823-833, 1991

[Hippisley-Cox et al. (2007)] Julia Hippisley-Cox, Carol Coupland, Yana Vinogradova, John Robson, Margaret May, and Peter Brindle. Derivation and validation of QRISK, a new cardiovascular disease risk score for the United Kingdom: prospective open cohort study. *BMJ*, 335 (7611): 136, Jul 2007.

[Ishwaran and Kogalur (2007)] H. Ishwaran and U.B. Kogalur. Random survival forests for R. *R News* 7(2):25—31, 2007

[Kothari et al. (2002)] Viti Kothari, Richard J Stevens, Amanda I Adler, Irene M Stratton, Susan E Manley, H. Andrew Neil, and Rudy R Holman. UKPDS 60: risk of stroke in type 2 diabetes estimated by the UK Prospective Diabetes Study risk engine. *Stroke*, 33 (7): 1776–1781, Jul 2002.

[Lagani and Tsamardinos (2010)] Vincenzo Lagani and Ioannis Tsamardinos. Structure-based variable selection for survival data. *Bioinformatics*, 26(15):1887-1894, 2010

[Lan et al. (1994)] Shu-Ping Lan, Christopher M. Ryan, Kenneth M. Adams, Igor Grant, Robert K. Heated, Lawrence I. Rand, Alan M. Jacobson, David M. Nathan and Patricia A. Cleary. A screening algorithm to identify clinically significant changes in neuropsychological functions in the diabetes control and complications trial. *Journal of Clinical and Experimental Neuropsychology*,16(2):303-316, 1994.

[Lauritzen and Spiegelhalter (1988)] S. L. Lauritzen and D. J. Spiegelhalter. Local computations with probabilities on graphical structures and their application to expert systems (with discussion). Journal of Royal Statistical Society, Series B, 50(2):157–224, 1998

[Skevofilakas et al. (2010)] Marios Skevofilakas, Konstantia Zarkogianni, Basil G Karamanos, and Konstantina S Nikita. A hybrid Decision Support System for the risk assessment of retinopathy development as a long term complication of Type 1 Diabetes Mellitus. *Conf Proc IEEE Eng Med Biol Soc*, 2010: 6713–6716, 2010. doi: 10.1109/IEMBS.2010.5626245. URL http://dx.doi.org/10.1109/-IEMBS.2010.5626245.

[Statnikov et al. (2005)] A. Statnikov, C. F. Aliferis, I. Tsamardinos, D. Hardin and S. Levy. A Comprehensive Evaluation of Multicategory Classification Methods for Microarray Gene Expression Cancer Diagnosis. Bioinformatics, 21(5):631-43, 2005

[Stevens et al. (2001)] R. J. Stevens, V. Kothari, A. I. Adler, I. M. Stratton, and United Kingdom Prospective Diabetes Study (UKPDS) Group. The UKPDS risk engine: a model for the risk of coronary heart disease in Type II diabetes (UKPDS 56). *Clin Sci (Lond)*, 101 (6): 671–679, Dec 2001.

[Stevens et al. (2004)] Richard J Stevens, Ruth L Coleman, Amanda I Adler, Irene M Stratton, David R Matthews, and Rury R Holman. Risk factors for myocardial infarction case fatality and stroke case fatality in type 2 diabetes: UKPDS 66. *Diabetes Care*, 27 (1): 201–207, Jan 2004.

[Tsamardinos et al. (2006)] Ioannis Tsamardinos, Laura E. Brown and Constantin F. Aliferis. The Max-Min Hill-Climbing Bayesian Network Structure Learning Algorithm. Machine Learning, 65(1):31-78, 2006

[UK Renal Association (2011)] UK Renal Association. Detection, Monitoring and Care of Patients with CKD. Available at http://www.renal.org/Clinical/GuidelinesSection/Detection-Monitoring-and-Care-of-Patients-with-CKD.aspx#downloads

[UKPDS Group (1991)] UKPDS Group. UK Prospective Diabetes Study (UKPDS). VIII. Study design, progress and performance. *Diabetologia*, 34 (12): 877–890, Dec 1991.

[Vergouwe et al. (2010)] Y. Vergouwe, S. S. Soedamah-Muthu, J. Zgibor, N. Chaturvedi, C. Forsblom, J. K. Snell-Bergeon, D. M. Maahs, P-H. Groop, M. Rewers, T. J. Orchard, J. H. Fuller, and K. G M Moons. Progression to microalbuminuria in type 1 diabetes: development and validation of a prediction rule. *Diabetologia*, 53 (2): 254–262, Feb 2010. doi: 10.1007/s00125-009-1585-3. URL http://-dx.doi.org/10.1007/s00125-009-1585-3.

[Wells et al. (2008)] Brian J Wells, Anil Jain, Susana Arrigain, Changhong Yu, Wayne A Rosenkrans, and Michael W Kattan. Predicting 6-year mortality risk in patients with type 2 diabetes. *Diabetes Care*, 31 (12): 2301–2306, Dec 2008. doi: 10.2337/dc08-1047. URL http://dx.doi.org/10.2337/dc08-1047.

Cohen AM, Hersh WR: A survey of current work in biomedical text mining, Briefings in Bioinformatics. Vol. 6 No. 1 57–71 March 2005

Shahar Y, Musen MA: Knowledge-based temporal abstraction in clinical domains, Artificial Intelligence in Medicine, Volume 8, Issue 3, July 1996, Pages 267-298

Spasic I, Ananiadou S, McNaught J and Kumar A: Text mining and ontologies in biomedicine: Making sense of raw text, Briefings in Bioinformatics. Vol. 6 No. 3 239–251 September 2005

Styltsvig HB: Ontology-based information retrieval, PhD thesis at Roskilde University, Denmark, 2006

Swanson DR: Undiscovered public knowledge. Libr. Q. 56(2):103-118. 1986

Vallet D, Fernández M, Castells P: An Ontology-Based Information Retrieval Model in Lecture Notes in Computer Science Volume 3532, 2005 103-110

Van der Zanden G: Quality Assessment of Medical Health Records using Information Extraction, Master Thesis at the University of Twente, August 19, 2010

Verma A: Ontology Based Personalized Modeling for Chronic Disease Risk Evaluation and Knowledge Discovery: An Integrated Approach, 2009

Villarreal V, Hervás R, Fdez AD, Bravo J: Applying ontologies in the development of patient mobile monitoring framework, 2nd International Conference on e-Health and Bioengineering - EHB 2009, 17-18th September, 2009, Iaşi-Constanţa, Romania

Yoo I, Song M: Biomedical Ontologies and Text Mining for Biomedicine and Healthcare: A Survey, Journal of Computing Science and Engineering, Vol. 2, No. 2, June 2008, Pages 109-136

Umpierrez, G E et al 2002, “Hyperglycaemia: An independent marker of in-hospital mortality in patients with undiagnosed diabetes”, J Clin Endocrinol Metab, vol.87, pp. 978-982.

Levatan, C S & Magee M F 2000, “Hospital management of diabetes”, Clin North Am, vol. 29, pp. 745-770.

Finney, S J et al 2003, “Glucose control and mortality in critically ill patients”, JAMA, vol. 290, pp. 2041-2047.

Dungen, K M et al 2009, ”Stress hyperglycaemia”, Lancet, vol. 373, pp. 1798-1807.

Clement, S et al 2004, ”Management of diabetes and hyperglycaemia in hospitals”, Diabetes Care, vol. 27, pp. 553-597.

Pompselli, J J et al 1998, “Early postoperative glucose control predicts nosocomial infection rate in diabetic patients”, JPEN J Parenter Enteral Nutr, vol. 22, pp. 77-81.

Van den Berghe, G et al 2001, ”Intensive insulin therapy in the critically ill patients”, N Engl J Med, vol. 345, pp. 1359-1367.

NICE Sugar Study Group 2009, “Intensive versus Conventional Glucose Control in Critically Ill Patients”, N Engl J Med, vol. 360, pp. 1283-1297.

Schnipper, J L et al 2006, “Inpatient management of diabetes and hyperglycaemia among general medicine patients at a large teaching hospital”, J Hosp Med, vol. 1, pp. 145-150.

Umpierrez, G E & Maynard, G 2006, “Glycemic chaos (not glycemic control) still the rule for inpatient care”, J Hosp Med, vol. 1, pp. 141-144.

Sakharova, O V & Inzucchi, S E 2005, "Treatment of diabetes in the elderly. Addressing its complexities in this high-risk group”, Postgrad.Med, vol. 118, pp. 19-26, 29.

Inzucchi, S E 2006, "Clinical practice. Management of hyperglycaemia in the hospital setting", N Engl J Med, vol. 355(18), pp. 1903-1911.

Clement, S et al 2004, “Management of Diabetes and Hyperglycemia in Hospitals”, Diabetes Care, vol. 27, pp. 553-591.

American Diabetes Association 2010, "Standards of medical care in diabetes - 2010", Diabetes Care, vol. 33, pp. 11-61.

Goldberg, P A et al 2006, "Glucometrics"--assessing the quality of inpatient glucose management“, Diabetes Technol Ther, vol. 8(5), pp. 560-9.

Umpierrez, G E et al 2009, “Comparison of inpatient insulin regimens with detemir plus aspart versus neutral protamine hagedorn plus regular in medical patients with type 2 diabetes”, J Clin Endocrinol Metab, vol. 94, pp. 564-569.

Umpierrez, G E et al 2007, “Randomized study of basal-bolus insulin therapy in the inpatient management of patients with type 2 diabetes (RABBIT 2 trail)”, Diabetes Care, vol. 30, pp. 2181-2186.

Umpierrez, G E et al 2011, “Randomized study of basal-bolus insulin therapy in the inpatient management of patients with type 2 diabetes undergoing general surgery (RABBIT 2 surgery)”, Diabetes Care, vol. 34, pp. 256-261.

Neubauer, K; Plank, J; Schaupp, L; Hoell, B; Spat, S; Beck, P; Buttinger, M; Schneeberger, M; Pieske, B; Pieber, T (2011) “Assessment of In-Hospital Glycaemic Management in Non-Critically Ill Patients International Hospital Diabetes Meeting”.; International Hospital Diabetes Meeting; Nov 17-19, Barcelona, SPAIN.

Andrews, R. C. and B. R. Walker (1999). "Glucocorticoids and insulin resistance: old hormones, new targets." Clin Sci (Lond) 96(5): 513-523.

Blakemore, A., S. H. Wang, et al. (2008). "Model-based insulin sensitivity as a sepsis diagnostic in critical care." J Diabetes Sci Technol 2(3): 468-477.

Boron, W. F. and E. L. Boulpaep (2008). Medical Physiology, Saunders.

Brannmark, C., R. Palmer, et al. (2010). "Mass and information feedbacks through receptor endocytosis govern insulin signaling as revealed using a parameter-free modeling framework." J Biol Chem 285(26): 20171-20179.

Buckingham, B., H. P. Chase, et al. (2010). "Prevention of nocturnal hypoglycemia using predictive alarm algorithms and insulin pump suspension." Diabetes Care 33(5): 1013-1017.

Cameron, F., G. Niemeyer, et al. (2008). "Statistical hypoglycemia prediction." J Diabetes Sci Technol 2(4): 612-621.

Care, B. R. and H. A. Soula (2011). "Impact of receptor clustering on ligand binding." BMC Syst Biol 5: 48.

Dalla Man, C., A. Caumo, et al. (2004). "Minimal model estimation of glucose absorption and insulin sensitivity from oral test: validation with a tracer method." Am J Physiol Endocrinol Metab 287(4): E637-643.

Dalla Man, C., R. A. Rizza, et al. (2007). "Meal simulation model of the glucose-insulin system." IEEE Trans Biomed Eng 54(10): 1740-1749.

Dassau, E., F. Cameron, et al. (2010). "Real-Time hypoglycemia prediction suite using continuous glucose monitoring: a safety net for the artificial pancreas." Diabetes Care 33(6): 1249-1254.

Duckworth, W. C., R. G. Bennett, et al. (1998). "Insulin degradation: progress and potential." Endocr Rev 19(5): 608-624.

Eren-Oruklu, M., A. Cinar, et al. (2010). "Hypoglycemia prediction with subject-specific recursive time-series models." J Diabetes Sci Technol 4(1): 25-33.

Hann, C. E., J. G. Chase, et al. (2005). "Integral-based parameter identification for long-term dynamic verification of a glucose-insulin system model." Comput Methods Programs Biomed 77(3): 259-270.

Hovorka, R., L. J. Chassin, et al. (2008). "A simulation model of glucose regulation in the critically ill." Physiol Meas 29(8): 959-978.

Hovorka, R., L. J. Chassin, et al. (2004). "Closing the loop: the adicol experience." Diabetes Technol Ther 6(3): 307-318.

Kiselyov, V. V., S. Versteyhe, et al. (2009). "Harmonic oscillator model of the insulin and IGF1 receptors' allosteric binding and activation." Mol Syst Biol 5: 243.

Kumar, S. and S. O'Rahilly (2005). Insulin Resistance, John Wiley & Sons.

Lin, J., D. Lee, et al. (2008). "Stochastic modelling of insulin sensitivity and adaptive glycemic control for critical care." Comput Methods Programs Biomed 89(2): 141-152.

Lin, J., J. D. Parente, et al. (2011). "Development of a model-based clinical sepsis biomarker for critically ill patients." Comput Methods Programs Biomed 102(2): 149-155.

Lin, J., N. N. Razak, et al. (2011). "A physiological Intensive Control Insulin-Nutrition-Glucose (ICING) model validated in critically ill patients." Comput Methods Programs Biomed 102(2): 192-205.

Palerm, C. C., J. P. Willis, et al. (2005). "Hypoglycemia prediction and detection using optimal estimation." Diabetes Technol Ther 7(1): 3-14.

Wanant, S. and M. J. Quon (2000). "Insulin receptor binding kinetics: modeling and simulation studies." J Theor Biol 205(3): 355-364.

Baker, C.F, Fillmore, C.J. and Lowe, J.B.: The Berkeley FrameNet project, In Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics-Volume, Association for Computational Linguistics, 1998.

Baumgartner, P., Burchardt, A.: Logic Programming Infrastructure for Inferences on FrameNet. In Logics in Artificial Intelligence, Springer, 2004

Burchardt, A., Erk, K. and Frank, A.: A WordNet detour to FrameNet. Sprachtechnologie, mobile Kommunikation und linguistische Resourcen, 8 (2005), 408–421.

Giuglea, A.M. and Moschitti, A.: Shallow semantic parsing based on FrameNet, VerbNet and PropBank. In Proceeding of the 2006 conference on ECAI 2006: 17th European Conference on Artificial Intelligence August 29– September 1, 2006, Riva del Garda, Italy, IOS Press, 2006.

Scheffczyk, J., Pease, A. and Ellsworth, M.: Linking framenet to the suggested upper merged ontology. In Proceeding of the 2006 conference on Formal Ontology in Information Systems: Proceedings of the Fourth International Conference (FOIS 2006), IOS Press, 2006.

McCarthy MI: Genomics, type 2 diabetes, and obesity, N Engl J Med. 363 (24): 2339-2350 2010

Pearson J, Bergenstal R: Fine-Tuning Control: Pattern Management Versus Supplementation, Diabetes Spectrum Volume 14, Number 2, 2001

9. Figures

10. Tables

Appendix A

Number	Name	Type	Description
1	GROUP	Character	Randomization Group
2	PHASE	Numeric	Phase Randomized (2,3)
3	RETBASE	Character	Retinopathy at Baseline (PRIM,SCND)
4	ADULT	Numeric	Adult >=18 (0=no/1=yes)
5	AGE	Numeric	Age (From Master) at entry
6	DURATION	Numeric	Total Months Duration of IDDM at entry
7	F002DATE	Character	Date Form 002 Completed
8	OBONSET	Numeric	Month and Year of IDDM Onset
9	PPDUR	Numeric	Post-Pubescent Duration (mos)
10	MARRIED	Numeric	Marital Status (0=NOT Married,1=Married)
11	OBMARRY	Numeric	Marital Status
12	PRIORHYP	Numeric	Past Hx of Severe Hypo(0=0,1=1-2,2= >=3)
13	OBC9	Numeric	Needed IV Glucose
14	HOLLSCOR	Numeric	Hollingshead(2 Factor)Social Class score
15	OBPATJOB	Numeric	Patient's Occupation
16	OBPATED	Numeric	Patient's Education
17	FAMIDDM	Numeric	Family History of IDDM (0=no/1=yes)
18	FAMNIDDM	Numeric	Family History of NIDDM (0=no/1=yes)
19	FAMHT	Numeric	Family History of HT (0=no/1=yes)
20	FAMMI	Numeric	Family History of MI (0=no/1=yes)
21	FAMEYE	Numeric	Fam Hist Eye Dz due to Diabetes (0/1)
22	IW	Numeric	Gender Specific Ideal Body Weight
23	PIDW	Numeric	% Ideal Body Weight (OBWEIGHT/IW)*100
24	BMI	Numeric	Body Mass Index (kg/m2)
25	OBSEX	Numeric	Sex
26	OBWEIGHT	Numeric	Weight (kg) at Baseline
27	OBHEIGHT	Numeric	Height (cm) at Baseline

28	MBP	Numeric	Mean Arterial Pressure
29	OBDBP1	Numeric	Diastolic BP (mm Hg, Form 002)
30	OBSBP1	Numeric	Systolic BP (mm Hg, Form 002)
31	INSULIN	Numeric	Total Insulin Dosage Units/Weight (kg)
32	OBGU7	Numeric	Kidney or Bladder Inf. w/ Antibiotics
33	SMOKES	Numeric	Smoking Status(1=never,2=ever,3=current)
34	OBSMOK1	Numeric	Ever Smoked Cigarettes
35	OBSMOK2	Numeric	Now Smokes Cigarettes
36	DRINKS	Numeric	Current Drinker (0=no,1=yes)
37	EXERCISE	Numeric	Level of Exericse (1=Strenuous...4=Mild)
38	OBRACE	Numeric	Race
39	OBDKAHSP	Numeric	Hospitalizations for DKA in Past Year
40	OBHYPHSP	Numeric	Hospitalizations for Hypog. in Past Yr
41	OBC8A	Numeric	Lost Consciousness without Seizure
42	OBC8B	Numeric	Lost Consciousness with Seizure
43	OBNEUR2	Numeric	Seizures
44	OBPSYCH5	Numeric	Suicide Attempt
45	OBPSYCH7	Numeric	Psychiatric Treatment
46	OHF3RE	Numeric	Visual Acuity Score-Right Eye (Form 008)
47	OHF3LE	Numeric	Visual Acuity Score-Left Eye (Form 008)
48	RETPAT	Numeric	Level of Retinopathy (1 - 4)
49	RET5	Numeric	Retinopathy Level -Baseline (DCCT Scale)
50	RETPAT00	Numeric	Retinopathy Severity Level
51	NEURODEF	Numeric	Presence of Clinical Neuropathy
52	RRV00	Numeric	ANS - RR Variation (x 1000)
53	VALS00	Numeric	ANS - Valsalva Ratio
54	FULLIQ	Numeric	Full Scale IQ
55	EDUCAT	Numeric	Mean Education (Years) - Form 013
56	CALORIES	Numeric	Calories (kcal)

57	BDHRATE	Numeric	ECG Heart Rate
58	BCVAL2	Numeric	T2-Glucose (ser, pre) mg/dl
59	BCVAL5	Numeric	Stimulated C-Peptide(pmol/ml, Form 023A)
60	AER	Numeric	Albuminuri (mg/24hr, Form 23b)
61	SCR	Numeric	Serum Creatinine (mg/dl, Form 23b)
62	CREAT	Numeric	Creatinine Clearance (ml/mn, Form 23b)
63	CHOL	Numeric	Cholesterol (serum,mg/dl, Form 023C)
64	TRG	Numeric	Triglycerides (serum,mg/dl, Form 023C)
65	HDL	Numeric	HDL Cholesterol (serum,mg/dl, Form 023C)
66	LDL	Numeric	LDL Cholesterol (serum,mg/dl, Form 023C)
67	WPMEAN	Numeric	Within-Profile Mean Blood Glucose(mg/dl)
68	BCVAL25A	Numeric	T25-BGP1 mg/dl
69	BCVAL26A	Numeric	T26-BGP2 mg/dl
70	BCVAL27A	Numeric	T27-BGP3 mg/dl
71	BCVAL28A	Numeric	T28-BGP4 mg/dl
72	BCVAL29A	Numeric	T29-BGP5 mg/dl
73	BCVAL30A	Numeric	T30-BGP6 mg/dl
74	BCVAL31A	Numeric	T31-BGP7 mg/dl
75	GFRX00	Numeric	Glomerular Filtration Rate (ml/min)
76	TSCGSI	Numeric	T-Score Global Severity Index
77	TSCDEP	Numeric	T-Score Depression
78	TOTQOL	Numeric	Quality of Life Total Score
79	HBA00	Numeric	Hemoglobin A1c at Baseline (Form 066)
80	HBAEL	Numeric	Hemoglobin A1c at Eligibility

Table 12: Baseline information employed for producing the long term risk assessment models

Appendix B

	Cox Regression	AFT	RSF	Ridge Cox Regression
No Selection	0.19	0.17	0.16	0.16
Univariate Selection	0.16	0.16	0.16	0.16
Forward Selection	0.17	0.16	0.16	0.16
Lasso Selection	0.17	0.17	0.17	0.17
BVS	0.18	0.16	0.16	0.16
SMMPC	0.16	0.16	0.16	0.16

Table 13: IBS results for the Hypoglycaemia outcome.

	Cox Regression	AFT	RSF	Ridge Cox Regression
No Selection	0.09	0.07	0.06	0.06
Univariate Selection	0.14	0.06	0.06	0.06
Forward Selection	0.51	0.07	0.06	0.06
Lasso Selection	0.06	0.06	0.06	0.06
BVS	0.07	0.07	0.06	0.06
SMMPC	0.06	0.06	0.06	0.06

Table 14: IBS results for the Ketoacidosis outcome.

	Cox Regression	AFT	RSF	Ridge Cox Regression
No Selection	0.23	0.16	0.16	0.16
Univariate Selection	0.16	0.16	0.16	0.16
Forward Selection	0.17	0.15	0.16	0.16
Lasso Selection	0.17	0.16	0.17	0.17
BVS	0.17	0.15	0.15	0.16
SMMPC	0.16	0.15	0.16	0.16

Table 15: IBS results for the Nephropathy outcome.

	Cox Regression	AFT	RSF	Ridge Cox Regression
No Selection	0.16	0.15	0.17	0.17
Univariate Selection	0.16	0.14	0.16	0.17
Forward Selection	0.17	0.14	0.16	0.17
Lasso Selection	0.19	0.17	0.18	0.18
BVS	0.16	0.15	0.16	0.17
SMMPC	0.16	0.14	0.16	0.16

Table 16: IBS results for the Neuropathy outcome.

	Cox Regression	AFT	RSF	Ridge Cox Regression
No Selection	0.32	0.24	0.24	0.23
Univariate Selection	0.25	0.2	0.24	0.24
Forward Selection	0.23	0.23	0.24	0.24
Lasso Selection	0.27	0.24	0.25	0.26
BVS	0.27	0.22	0.24	0.24
SMMPC	0.24	0.2	0.24	0.23

Table 17: IBS results for the Neurobehavioral outcome.

	Cox Regression	AFT	RSF	Ridge Cox Regression
No Selection	0.3	0.11	0.11	0.11
Univariate Selection	0.12	0.11	0.1	0.11
Forward Selection	0.12	0.11	0.1	0.11
Lasso Selection	0.12	0.12	0.12	0.12
BVS	0.11	0.11	0.1	0.11
SMMPC	0.12	0.11	0.1	0.11

Table 18: IBS results for the Retinopathy outcome.