

Date of publication xxxx 00, 0000, date of current version xxxx 00, 0000.

Digital Object Identifier 10.1109/ACCESS.2017.Doi Number

A Data Analytics Suite for Exploratory Predictive, and Visual Analysis of Type 2 Diabetes

NADA Y. PHILIP¹ Senior Member, IEEE, MANZOOR RAZAAK¹, JOHN CHANG², MAURICE O’KANE³, SUCHETHA M⁴ and BARBARA K PIERSCIONEK⁵

¹Digital Media for Health Research Group, School of Computer Science and Mathematics, Kingston University London, Surrey, UK

²National Health Service (NHS), Croydon, UK

³Clinical Chemistry Laboratory, Altnagelvin Hospital, Western Health and Social Care Trust, Londonderry, N. Ireland

⁴Division of Healthcare Advancement, Innovation and Research, Vellore Institute of Technology, Chennai, India

⁵Faculty of Health Education Medicine and Social Care, Medical Technology Research Centre, Anglia Ruskin University, UK

Corresponding author: Nada Y. Philip (e-mail: n.philip@kingston.ac.uk).

The research leading to these results have received funding from the European Commission funded program under H2020 Grant Agreement No: 644906 (AEGLE), <http://www.aegle-uhealth.eu>

ABSTRACT Long-term management of chronic disorders such as Type 2 Diabetes (T2D) requires personalised care for patients due to variation in patient characteristics and their response to a specific line of treatment. The availability of large volumes of electronic records of T2D patient data provides opportunities for application of big data analysis to gain insights into the disease manifestation and its impact on patients. Data science in healthcare has the potential to identify hidden knowledge from the database, re-confirm existing knowledge, and aid in personalising treatment. In this paper, we present a suite of data analytics for T2D disease management that allows clinicians and researchers to identify associations between different patient biological markers and T2D related complications. The analytics suite consists of exploratory, predictive, and visual analytics with capabilities including multi-tier classification of T2D patient profiles that associate them to specific conditions, T2D related complication risk prediction, and prediction of patient response to a particular line of treatment. The analytics presented in this paper explore advanced data analysis techniques, which are potential tools for clinicians in decision-making that can contribute to better management of T2D.

INDEX TERMS Big data for healthcare, data analytics, personalized care, healthcare data visualisation, prediction analytics, risk prediction, T2D.

I. INTRODUCTION

The rapid technological advancements in cloud technologies, big data infrastructure, and artificial intelligence have generated significant excitement in developing data-driven solutions for various domains including the healthcare sector. Development of big data infrastructure, including data analytics for healthcare applications requires careful design and planning supported by close collaboration between healthcare experts and relevant stakeholders due to the sensitive nature of the healthcare data involved and the impact it may have on patients’ well-being.

The project AEGLE, commissioned by European Union

(EU), developed a big data framework aimed at providing big data services for healthcare, including electronic healthcare record data storage, data analytics, cloud services for accelerated training of complex analytics, and real-time processing of large data volumes. Figure 1 provides an overview of the AEGLE ecosystem. Further details of the AEGLE system can be found at [1]. Under the AEGLE project a host of data analytics was developed including analytics for Type 2 Diabetes (T2D) amongst others.

T2D is a chronic condition with increasing prevalence across the globe. T2D is one of the common causes of

morbidity and mortality, leading to significant consumption (PHE) reported that 3.8 million people in England aged over 16 had diabetes [2], and it is estimated to have leading cause of death globally [4]. The impact of diabetes on economic costs is significant and was estimated to be \$327 billion in the USA alone [5]. Therefore, early intervention and effective treatment strategies are necessary to reduce T2D impact on patient quality of life and economic costs.

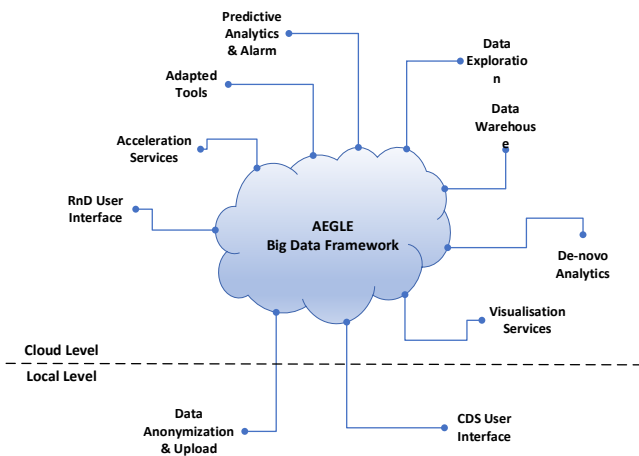


FIGURE 1. Overview of AEGLE big data framework.

The increase in electronic recording of patient data over the last few decades provide data analysts opportunities for exploring healthcare databases to identify previously unknown patterns and associations that are potentially useful for a better understanding of disease and their management. Historical data of the patient cohort enables analysts to develop analytics to predict patient disease progression and personalise treatment strategy accordingly [6].

Data analytics for T2D is a well explored topic and substantial amount of research papers can be found in the literature. One of the most widely explored topic in T2D is complication risk prediction. Various models exist ranging from the classical Cox's models and its variations [7] [8] [9] [10] to the more recent machine learning based models based on methods including support vector machine (SVM) [11], naives bayes [12], nearest neighbor [13], random forest [14], logistic regression [15], genetic algorithm [16] and deep learning [17] [18] [19].

The progress in development of data analytics for T2D provides a potential for development of a tool that empowers healthcare professionals in data analysis and decision making. In this paper, we present our work on data analysis of T2D data designed for finding associations between different patient markers, risk predictions for various complications, and prediction of patient response to medications. The individual analytics developed are

of healthcare resources. In 2015, Public health England increased to 4.7 million people in 2019 [3]. World Health Organization (WHO) estimates T2D to be the seventh presented as a first step towards a T2D analytics suite with a goal that the suite enables clinicians and researchers to gain insights into T2D disease and its management.

The intention of the analytics suite proposition is to **emphasise the** potential of using a host of data analytics as a toolbox by healthcare stakeholders for patient data analysis and decision making. An initial assessment of the feasibility of the presented analytics suite was performed as part of the AEGLE project's cloud based big data platform for healthcare data analysis [1] [20]. Initial assessments on the T2D analytics suite were performed by clinicians and received positive feedback.

The analytics presented in this paper are not limited in terms of novelty and clinical significance. The analytics suite represents initial steps towards developing a framework for data analytics suite for clinical practice that will provide a novel and much needed diagnostic tool. Clinicians treating T2D do not always have sufficient information from presenting signs or symptoms to know definitively which medication or course of treatment will work. The vast variability in T2D patients and their characteristic features that can influence the course of the disease and the response to treatment makes it difficult to know which type of treatment may be best for which patient without testing certain medications to gauge response. A tool that will cohort patients with similar risk factors for T2D will provide greatly improved indicators for what course of treatment is best, minimising side-effects and optimising treatment outcomes with a personalised approach. The data analytics platform described in this paper has the potential to offer such a tool to clinicians.

To the best of our knowledge, our paper is one of the earliest attempts towards the development of framework for a data analytics suite for T2D.

The rest of the paper is organised as follows. In Section II an introduction to the T2D analytics suite is given. Section III describes the patient profile classifier analytic workflow. The risk prediction analytics is described in Section IV, followed by description of response prediction analytics in Section V. Section VI discusses the challenges in healthcare data analysis and Section VII concludes the paper.

II. ANALYTICS SUITE FOR T2D DATA ANALYSIS

The data analytics developed for T2D is presented in this section. Figure 2 illustrates the methodology followed for the development of the proposed analytics. The methodology shown is a typical approach followed in data analytics development, however, close collaboration with clinicians at all stages and specifically at the requirements gathering, analytic approach, data collection, modelling and

feedback stages is crucial in healthcare related data analysis process.

The T2D analytics suite aims to develop exploratory, predictive, and visual analytics. The exploratory analytics focus on exploring the diabetes dataset for performing operations such as raw data pre-processing, classification, and associations between different patient markers and diabetes-related complications, and hypothesis generation. The predictive analytics are concerned with identifying the risk carried by diabetes patients for a complication based on the patient biological markers and the probability of the patient developing a complication over time in future. The analytics also focuses on predicting the response of diabetes patients to a particular line and combination of treatments. The visual analytics include custom visualisations developed for T2D data analysis that allows clinicians to gain a perceptible insight into the disease and impact on patients.

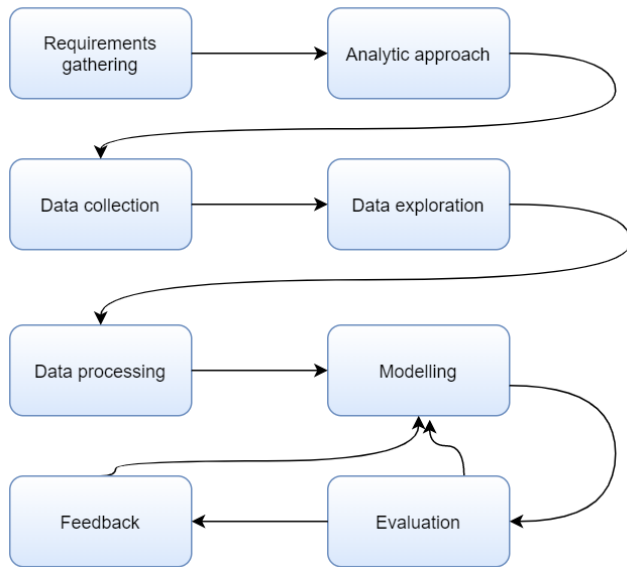


FIGURE 2. Methodology followed for the development of the T2D data analytics.

T2D patient data were mainly obtained from two data sources: Croydon/Prowellness database and Diamond database. A summary of the two datasets is given in Table I. From the two databases, patient biological markers that are risk indicators of diabetes related complications capable of providing insights into patient response to treatments were identified. After extensive consultations with healthcare experts and with the aid of big data analysis techniques, several analytics were developed.

The principle approach followed for the T2D data analysis was to cluster the patients according to their demographics and biological markers and investigate their associations with known T2D related complications

followed by the development of predictive analytics to model the associations between different patient markers. This approach helped to gain insight into hypothesis building. For instance, the example heatmap in Figure 3 obtained on a synthetic dataset illustrates how associations between different variables in a dataset can be explored.

The exploratory part of the heatmap analytic identifies associations between the chosen variables within the T2D database and utilises multiple statistical analysis, including correlation analysis and chi-squared analysis. The findings are then visualised on a heatmap wherein the associations between the variables are represented by means of dendrograms.

In Figure 3 heatmap, the correlation between patient biological markers listed on the y-axis and three risk conditions: visual impairments, renal replacement therapy, and death is presented. In the heatmap, light shades of blue corresponds to strong correlation between the marker variables and the complication risk and darker shades of blue correspond to weak correlations.

The clinicians and researchers can use this analytic to understand better associations between variables or confirm already known strong and weak associations between different variables and thus generate hypotheses for further research and analytic development.

Three T2D analytics workflows were developed, namely: patient profile classifier, complication risk predictor, and patient treatment response predictor. Each T2D analytics workflow further includes multiple analytics. Details on the three analytics workflows, including its development, implementation, and findings are provided in the next three sections respectively.

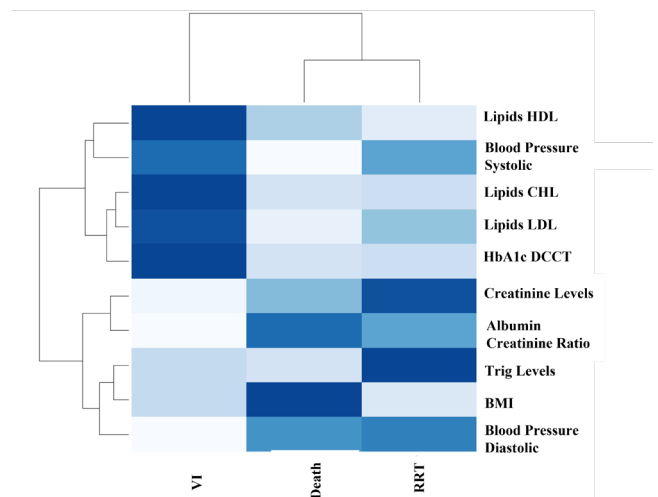


FIGURE 3. Heatmap of correlation of patient marker variables (on y-axis) with three risk conditions (on x-axis). VI = Visual Impairment, RRT = Renal Replacement Therapy. Light colours indicate relatively strong correlations between the variable and the risk condition.

TABLE I
SUMMARY OF THE DATASETS USED FOR DEVELOPING T2D DATA ANALYTICS

	Croydon/Prowellness	Diamond
Type	Combination of database, commercial, and NHS, on secondary care of people with diabetes in South-West London/Surrey	Clean structured, commercially curated database on tertiary care of people with diabetes in Northern Ireland
Description	Parameters related to the patients and their diabetes. Longitudinal records of the patients' diabetes over several years, progression of the disease	demographics, medications, anthropomorphic markers, social and demographic factors, biochemical markers, lifestyle factors, lab results, clinical notes
Quantity	19,186 patients	16,936 patients
Mean Age (years)	60.85	61.65
Men	10432	9937
Women	8354	7155
Values at baseline:		
HbA1c recorded	12,358	10,884
Mean HbA1c (mmol/ml)	71.27	67.25
BMI recorded	17,974	15,586
Mean BMI	30.30	31.40
B.P. (systolic) recorded	15,895	12,664
Mean B.P. (systolic)	134.95	136.86

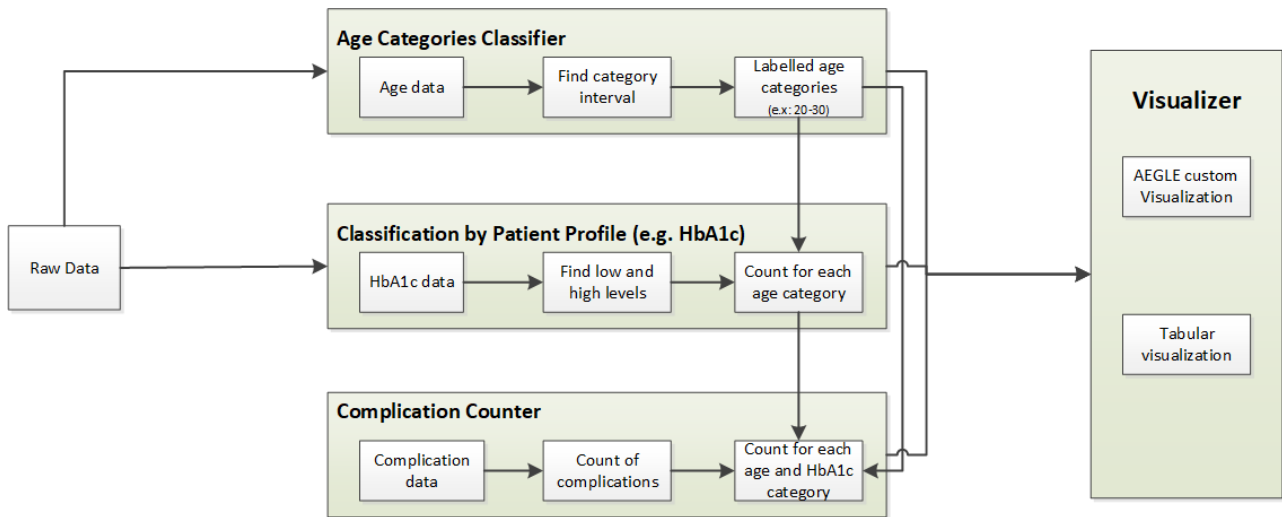


FIGURE 4. Patient Profile Classifier Workflow. A two-tier classification is performed by the analytics to associate patient demographics with patient markers and T2D related complication.

III. APATIENT PROFILE CLASSIFIER

The patient profile classifier workflow includes both exploratory and visual analytics. Figure 4 illustrates the components of the workflow. The classifier aims to classify patients according to a preferred demographic category (e.g., age or gender) and a biological marker class (e.g., HbA1c levels) and associate the classes to a complication (e.g., blindness). This multi-tier classification is achieved where the patient profile is further classified according to patient biological marker levels such as HbA1c, Lipids, etc., and the count of patients for each age category and marker class is obtained. The threshold for the marker levels can be set to a chosen value by the user. Further, under each patient marker class, the count of patients associated with complications such as amputation, visual impairments, etc., is fetched.

by means of a population pyramid analytic that helps to understand the composition of the population according to chosen criteria [21]. Further, a custom visualisation analytic is built based on the population pyramid analysis.

A. POPULATION PYRAMID ANALYTIC

In Figure 4, the exploratory part of the analytic performs a multi-step classification of the diabetes population starting with age categories. Next, in each age category, the patient

The data obtained from the exploratory analytic provides a two-tiered population pyramid distribution of the T2D patients, i.e., based on a chosen demographic category and biological marker. To visualise the two-tier population pyramid, an interactive, custom visualisation analytic was developed that plots the findings of the exploratory analytic as a stacked bar-population pyramid graph. The exploratory

and visualisation analytic was implemented in R programming language and using the rCharts [22], HighCharts [23], visualisation packages. Figure 5 shows the stacked-bar population pyramid chart where the T2D patient data is categorized into age categories, low and high HbA1c levels and then associated with patients suffering from visual impairment complications. The visualisation helps clinicians to get an overview of the prevalence of a T2D related complication in a patient subgroup from a database.

III. T2D PATIENT COMPLICATION RISK PREDICTOR

T2D patients are associated with increased risk for complications such as visual impairment, coronary heart disease, amputations, renal impairment, or stroke [24]. Risk prediction models are beneficial for clinicians and patients to understand their likeliness of developing a complication based on their current T2D condition. Risk calculators such as QRISK2 are recommended for identifying cardiovascular risk among T2D patients [25]. Risk estimation for a complication is done by determining the factors that are likely to cause a complication risk. Identification of potential risk factors and timely intervention for control and management of the risk factors can reduce the patient's risk for T2D related complications [26]. A complication risk prediction model based on the Cox's proportional hazards model is presented in this section.

A. PREDICTION ANALYTICS DEVELOPMENT

Risk prediction models for several complications including visual impairment, toe amputation, stroke, cardiovascular risk, and renal impairments, were developed. A cohort study of the patient data from the Croydon/Prowellness database was conducted. Through consultation with clinicians and literature research, several predictor variables with established risk factors for T2D related complications were identified and included the variables: age, HbA1c values, blood pressure, BMI, and lipids [7].

The dataset consisted missing values in all of the variables. To addressing missing values, multiple imputations based on Rubin's rules [27] and available in the R MICE [28] package was applied. The imputation over ten iterations were applied on the database to estimate values for missing values in the database and for better complete case analysis [29]. Data standardisation was performed by conversion of metric units (e.g., mg/dL to mmol/L) and merging of similar data columns.

For risk prediction of a complication, the widely popular Cox's proportional hazards model (CPH) [30] was utilized to estimate the risk factors for each predictor variable. A censoring indicator is used to censor patients at the date of diagnosis of complications. Hence, the censoring indicator is considered to be either a complication or no complication. The survival time of the patients is computed for the patients from the time of diabetes diagnosis to the occurrence of the complication.

B. PREDICTION ANALYTIC RESULTS AND VALIDATION

TABLE II. HAZARD RATIOS FOR PREDICTOR VARIABLES OBTAINED FROM CPH MODEL FOR VARIOUS COMPLICATIONS.

Based on the pre-processed predictor variables, the CPH model computes the hazard ratios for the various predictor variables. The hazard ratios or the risk ratios indicate the extent of the risk carried by different predictor variables for a complication. Table II shows the hazard ratios of five predictor variables obtained for vision impairment and toe amputation risks from CPH models. The hazard ratios (HR) for a predictor variable are interpreted as follows:

- HR = 1: No effect on the complication
- HR > 1: Increases risk for complication
- HR < 1: Reduces risk for complication

The risk prediction models are validated using the 10-fold cross-Validation method. The Croydon database was

Predictor variables	Hazard ratios	
	Vision impairment	Amputation
HbA1c	1.15	1.69
Blood pressure	0.99	1.52
Age	0.97	1.03
BMI	1.13	1.01
Lipids	1.20	0.53

segregated to ten folds of training and test dataset. The risk models for each complication were validated separately using metrics such as sensitivity, specificity, and accuracy. The risk prediction model and the 10-fold cross-validation study was implemented in R programming environment. The sensitivity and specificity analysis for each fold of training and test dataset were analysed and the mean of the metrics across the ten folds was calculated. The sensitivity and specificity analysis follows the True Positive Rate and False Positive Rate estimation described in [31] for risk prediction survival models. The analysis is performed over a 15-year timeline and uses the *survAUC* R package [32]. The *sensitivity* and *specificity* values are used to compute the prediction accuracy of the model via $Accuracy = (sensitivity)(prevalence) + (specificity)(1 - prevalence)$, where the *prevalence* is the number of positive conditions over the total population. The outcomes of the results are presented in Table III.

C. VISUAL ANALYTIC FOR SURVIVAL PROBABILITY ANALYSIS

The CPH model allows the clinicians to obtain hazard scores for the predictor variables and get a global view on their impact on the risk for a complication. It would be beneficial for clinicians and patients to obtain a risk score for individual patients based on their biological markers. A survival

analysis based visualisation analytic is designed using the CPH model developed for complication risk predictions. Two sets of survival analysis curves are presented where (1) provides a global view of a predictor variable impact on a complication risk, and (2) provides a survival probability curve for individual patients based on their conditions.

In Figure 6, various survival probability curves for each complication obtained from the visual analytic are shown. Each probability curve for a complication demonstrates the

increase in risk for the complication over time for all the patients considered in the Croydon/Prowellness database. These survival curves enable clinicians and researchers to analyse complication risks for a cohort group of T2D patients from a dataset or geographical region and frame interventions and policies to reduce the risks of the complication. More insightful survival curves are obtained by demonstrating the impact of specific patient markers on the survival probability.

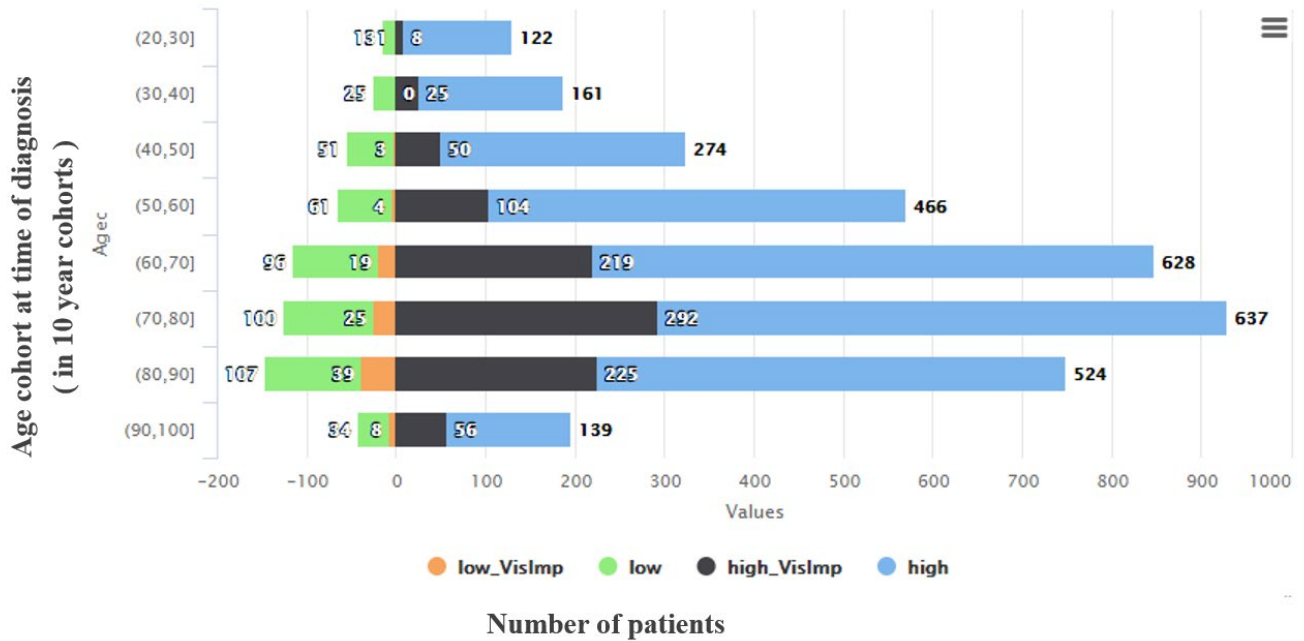


FIGURE 5. Population pyramid classification. Number of patients for each category is shown. Blue corresponds to patients in the database with high HbA1c values (i.e. above 7 mmol/L) and green corresponds to low HbA1c values. Black and Orange categories correspond to number of patients with visual impairment in high and low HbA1c classes respectively.

TABLE III. CPH RISK PREDICTION MODEL VALIDATION RESULTS FOR COMPLICATIONS.

Complication	Data volume	Accuracy (%)	Sensitivity (%)	Specificity (%)
Visual impairment	1220 patients	82.13	71.05	67.41
Amputation	91 patients	68.90	61.15	64.41
Cardio-vascular	144 patients	54.36	67.4	54
Renal impairment	186 patients	63.29	67	51.2

In Figure 7, the impact of four patient markers; HbA1c levels, body mass index (BMI), blood pressure, and high-density lipoprotein (HDL) lipids levels; on the risk for visual impairment is demonstrated. Each patient marker is categorized into low, medium, and high levels along with its

impact on the survival probability is shown in each individual sub-figure. For instance, in Figure 7c, the survival probability curve (in red) is higher for low BMI and the probability of survival decreases for medium BMI (green curve), and further decreases for high BMI (blue curve).

Conversely, in Figure 7d for HDL lipids where high level (in mmol/L) is considered healthy, the survival probability curves progressively increases for low, medium, and high HDL lipid levels.

In terms of an individual T2D patient, understanding their current risk for a complication can be an important information for the treatment and management of diabetes for clinician and the patient. Based on the CPH complication risk models, a survival analysis user interface was designed and developed for potential use by clinicians and patients.

Figure 8 shows the user interface for a complication risk analysis for patients. The user interface allows a clinician or patient to input their biological marker levels such as BMI, HbA1C, blood pressure, and other entries and obtain a survival probability curve for their current condition to understand their risk levels. This information can help clinicians and patients to formulate a strategy for the patients' diabetes management, for instance, lowering BMI or blood pressure control.

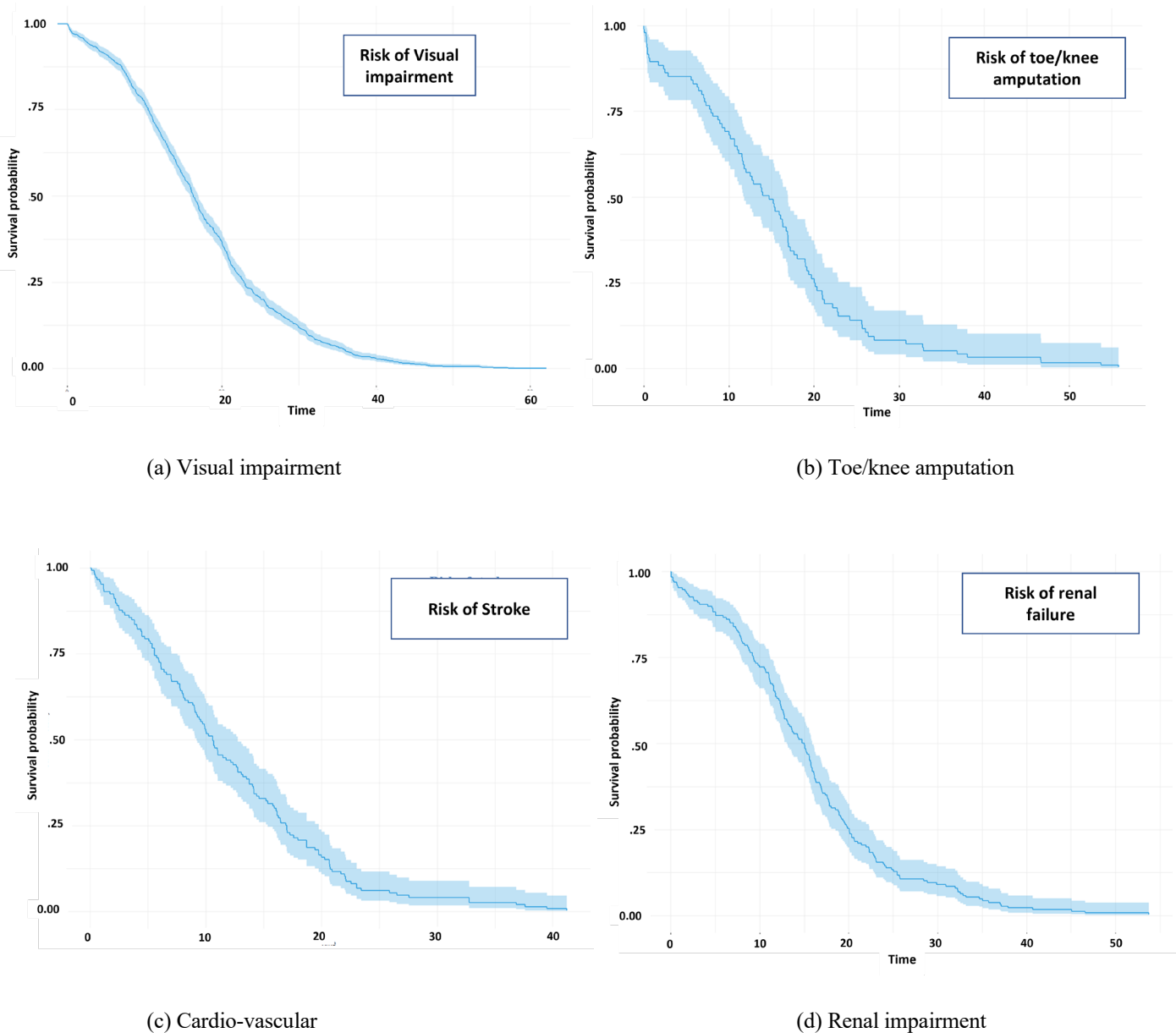


FIGURE 6. Survival probability curves show the rate of increase in risk for complication with time. Y-axis indicates the survival probability and x-axis is duration in years.

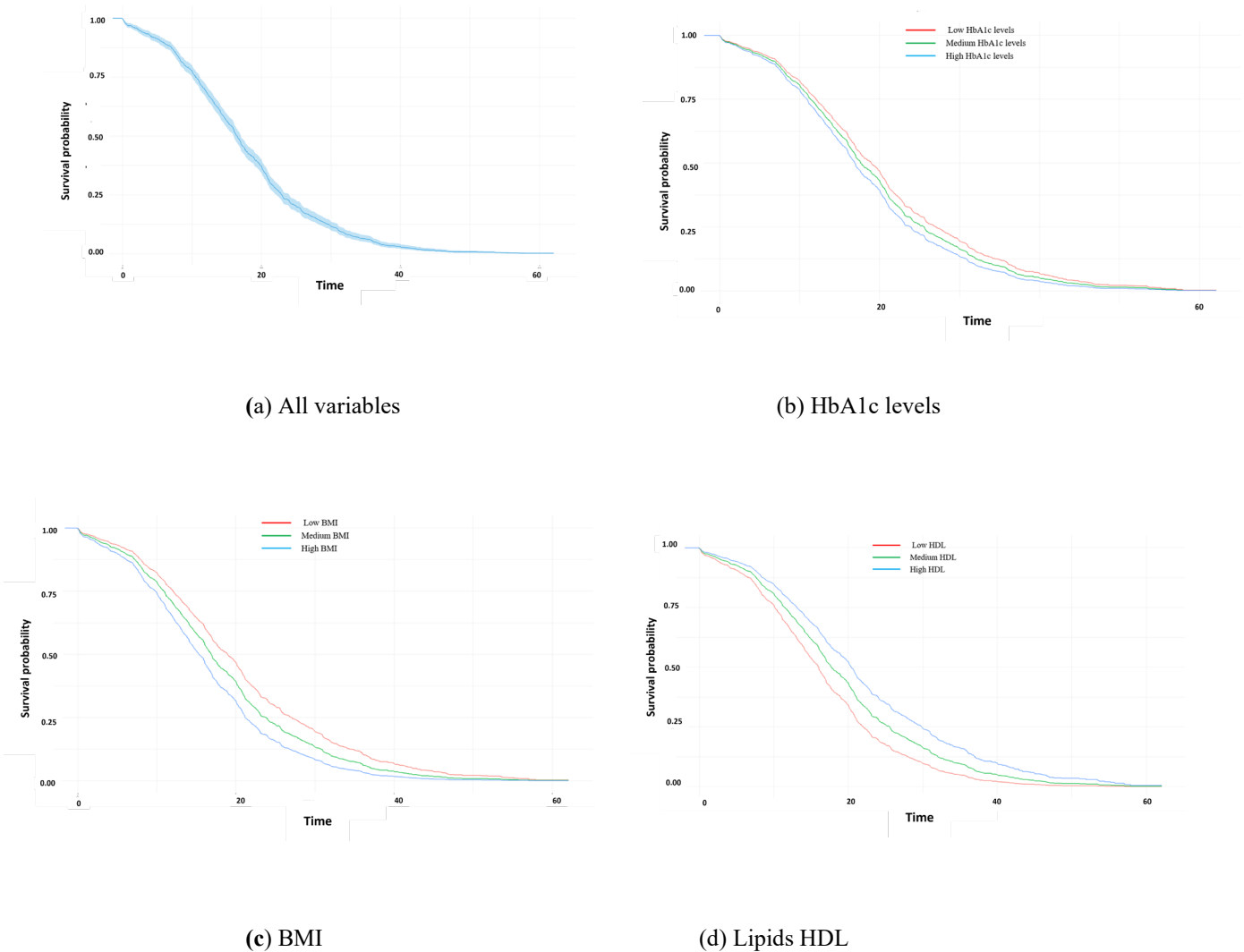


FIGURE 7. Survival probability curves for visual impairment illustrating the impact of predictor variables on the rate of risk. Y-axis indicates the survival probability and x-axis is duration in years. (a) shows survival probability curve when risk of all patient biomarkers are considered. (b) survival curve for low, medium and high levels of HbA1c (c) survival curve for low, medium, and high BMI levels (d) survival curves for low, medium, and high HDL levels. It can be observed that the survival curve generated by the analytic is variant and corresponds to the impact of the patient biomarker on the risk of visual impairment.

III. PATIENT TREATMENT RESPONSE PREDICTION

T2D is a complicated disorder and often correlates with several microvascular complications such as renal damage, retinopathy, and neuropathy along with other major complications such as stroke and heart disease [33]. Eleven distinct mechanisms are currently thought to cause diabetes, and over a hundred genetic points are associated with T2D. Based on various observed characteristics, different subgroups of diabetes patients are known [34].

T2D patients belonging to different subgroups tend to respond differently to different classes of medications. For instance, a combination of sodium-glucose transporter 2 inhibitors (SGLT2is) with glucagon-like peptide-1 (GLP-1) mimetics are common treatment drugs for T2D patients. However, people with type 2 diabetes who are insulinopaenic and therefore at risk of ketoacidosis [i.e. a build up of ketoacids in blood, a serious acute metabolic disturbance] should not be used in patients who are ketosis prone [20].

Identification of patient-related factors that potentially lead to non-response to specific treatments can help in

choosing the optimal class of treatments for a patient and thus personalise and accelerate the patient treatment with known effective medications. Further benefits include reducing the number of medication trials, cost savings, and decrease patient exposure to potential side effects.

Big Data analysis provides opportunities in identifying groups of patients who are likely to respond to a specific line of treatment. Patient cohort data over an extended period and a combination of dataset sources enables to detect patterns and patient response to medications over time. Analysis of cohort data provides the potential to build prediction models to predict patient response to specific medications based on patient characteristics. The Diamond database offers a cohort of T2D patients, including medications recommended to them. A machine learning-based prediction model is built for predicting patient response to third line agents (i.e., SGLT2is inhibitors). The SGLT2is inhibitors are used to promote glycosuria and are an approved class of drugs in the treatment of T2D.

A. TREATMENT RESPONSE PREDICTOR IMPLEMENTATION AND RESULTS

Before building the prediction model, it is necessary to train the machine learning model to learn to predict patient response to SGLT treatment. From the T2D Diamond database, patients treated with SGLT are considered to show good response when they show improvement in HbA1c levels by 11 mmol/L over three or six month period. Based on the HbA1c changes, the patient response is classified into the top and lower quantiles of good response and bad response category. In the next step, patient features are extracted that are known to potentially influence patient response to medications. The features selected include age, gender, duration of diabetes, weight, BMI, HbA1c levels, and medications taken.

A cohort data of approximately 2300 patients from the Diamond database was selected to train and build the prediction model. Out of the 2300 patients, approximately 1000 patients belonged to the good response category and the remaining to the bad response category. The dataset was further classified into a training dataset (80%) and a validation dataset (20%). The well-known machine learning model support vector machine (SVM) was chosen as our prediction model due to its proven high-performance accuracy in data prediction [35]. The prediction model was implemented on R programming environment. The SVM model was validated via a 5-fold Cross-Validation study. An average prediction accuracy of 65.05% was obtained in the 5-fold Cross-Validation, and the best prediction accuracy obtained was 73.3%. The outcomes of the prediction model are shown in Table IV.

TABLE IV. SUMMARY OF SVM PREDICTION MODEL FOR PATIENT RESPONSE TO SGLT LINE OF TREATMENT.

	Outcome
Patients (n)	2300
Validation method	5-fold cross-validation
Best accuracy	73.30%
Sensitivity	66.85%
Specificity	75.21%
Average accuracy	65.05%

The treatment response prediction model provides capabilities to analyse the response of patient subgroups to a specific line of treatments. This leads to a reduction in medication trials to find the most effective treatment for T2D patients. Inclusion of such analytics is beneficial for clinicians to predict the response of a patient of a certain profile to a particular line of treatment. With increasing data collection on patient response to medications, the presented prediction model can be further periodically trained to be more reliable and to give accurate predictions. In addition to models such as SVM, the presented approach can be adapted to include other machine learning models such as Naïves Bayes and k-nearest neighbor (kNN) for better support to clinicians with treatment decisions.

VI. DISCUSSION - CHALLENGES IN MEDICAL DATA ANALYSIS

The process of developing data analytics for healthcare presents unique challenges due to the sensitivity of the data involved and the impact of the outcomes. Some of the common challenges in data analysis along with recommendations to overcome the challenges are discussed below.

- **Data quality:** Like most big data solutions, data quality is a problem too in the healthcare domain. Main issues with data quality arise due to lack of completeness in the data (missing values), data repetition, irregular and inconsistent data update, less accurate and invalid data entries. Data pre-processing approaches such as multiple imputations and standardization of multiple data sources can address data quality issues. In our studies, several approaches to improve data quality were used. For instance, multiple imputations were part of data pre-processing for complication risk predictors presented in Section IV. The Diamond database and Prowellness databases were often merged to provide sufficient data volume for our analytics. Data standardisation was approached by conversion of metric units (e.g., mg/dL to mmol/L) and merging of similar data columns.

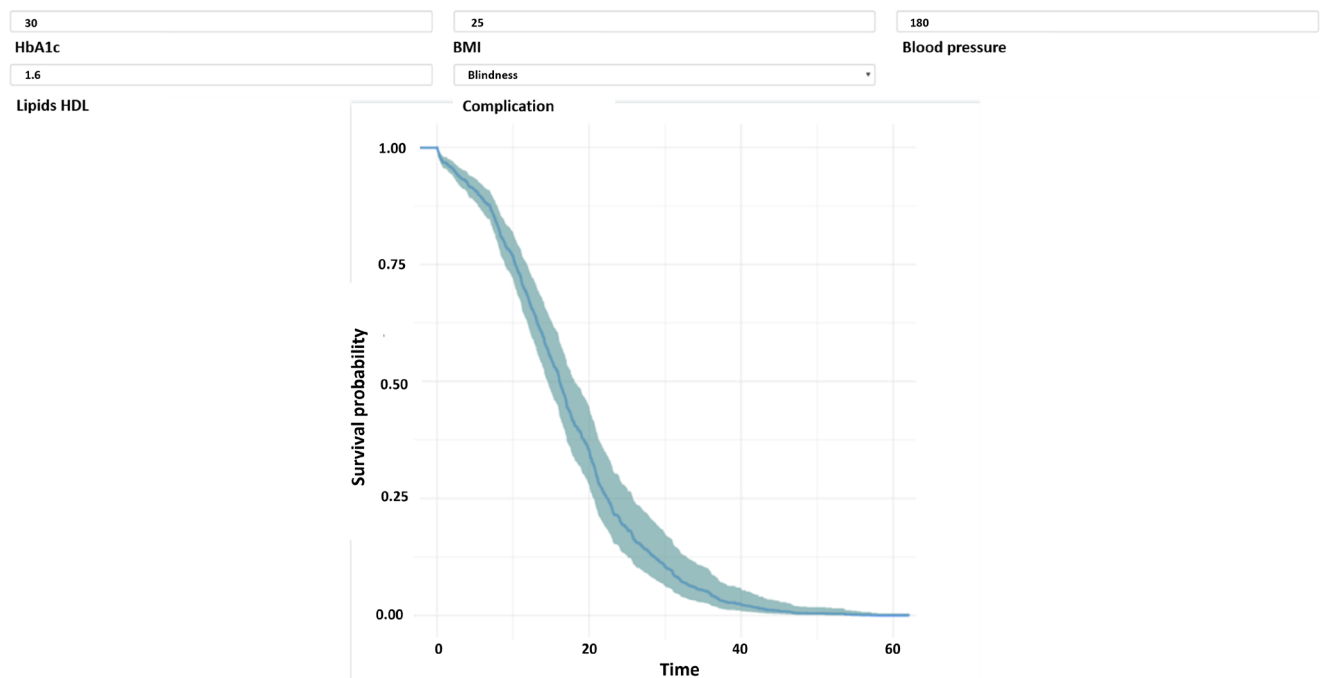


FIGURE 8. Interactive survival probability curve for a patient complication risk assessment. A clinician enters patient markers such as HbA1c, BMI and selects a complication and obtains a patient specific survival probability curve for the chosen complication.

- **Feasibility:** Developing big data analytics for healthcare requires multidisciplinary effort between data analysts and clinical experts. Prior to development of data analytic solutions, feasibility analysis is required. From the clinician's perspective, use case scenarios must be developed to describe the relevant clinical problem to be addressed. Clinicians must also analyse the data feasibility, i.e., availability and quality of data. For instance, during the course of our study, development of advanced deep learning based data analytic solutions though desired were not feasible due to lack of availability of large datasets. From data analysts perspective, technical feasibility analysts would allow the stakeholders to arrive at a feasible solution for potential development. Factors such as cost-effectiveness, clinical relevance, and application of outcomes in clinical trials must be evaluated before investing in development of data analytics.
- **Solutions for Decision Support System:** It is essential for the solutions developed to enact as a Decision Support System (DSS). A DSS would be an empowering tool for clinicians to make patient treatment-related decisions. For instance, the survival probability analyser UI presented in Section IV can be utilised by clinicians to understand a particular patient's present risks to a complication and use the knowledge to strategise on disease management. Further modifications to the probability analyser to enable it to describe a patient risk in descriptive form (e.g., low, medium, and high risk) can potentially be used by a patient to assess their risk category either independently or under guidance from a clinician.
- **Simplicity:** Often, solutions for many data analytics do not require Big Data infrastructure and can be solved with available data mining tools. However, as the volume and veracity of healthcare records are increasing, sophisticated big data tools are required. Despite the sophistication and complication required for the development of big data analytics, it is desirable for the analytics outcome to be a simplified solution capable of utilisation in daily clinical practice. One of the objectives of the various analytics presented in this paper is to present simplified solutions for understanding and improving T2D disease management. We believe the data analytics presented in the paper with further clinical validation have the potential to be adopted in clinical practice especially for activities such as data visualization and risk prediction. A key factor being that it does not require significant IT infrastructure and can be adapted according to the data availability and requirements of the clinician.
- **Extensibility:** A desirable feature for data analytics is to permit inclusion of additional features and models in analysis. The analytics presented in the paper can be extended to include relevant features for analysis. For instance, the population pyramid analyser and complication risk predictor can be extended to include

new biological markers for analysis and are not restricted to the markers presented in the paper. As discussed in Section V-A, the presented prediction model approach is not limited to SVM model alone and can be extended for use with other machine learning models.

- Scalability: A key characteristic for analytics is the ability to scale the analysis to larger datasets. The presented analytics are tested on relatively smaller datasets with lower than 20,000 patients. However, further tests are required to evaluate the capability for analysing significantly larger datasets.

The tool requires further testing of the analytics on a large scale using more external T2D databases. Planned future works include design of a robust framework for the analytics suite that includes flexibility for clinicians to choose from multiple models. Further, the data models in the analytics will be extended to include more advanced, clinically validated models.

VII. CONCLUSION

In this paper, we presented an analytics suite that performs exploratory, predictive, and visual analysis of T2D data. Three types of analytics workflows were presented that perform: (1) classification of T2D patients into required categories and identifying associations to a condition of interest, (2) analysis of T2D database to build a predictive model that can assess risk of patients to T2D related complications, and (3) prediction of patients' response to a specific line of treatment plan. The visual analytics provides a simplified representation of the outcome for clinicians and patients.

The analytics presented have the potential to support clinicians to decide treatment plans for T2D patients. This offers huge advantage that had not been previously possible for a more personalised approach to treating T2D that will be safer and more beneficial for the patient as it will minimise side effects and offer faster, more effective treatment. It will also provide economic advantages to the healthcare system.

Possibilities for future work include building and training the model on larger databases to increase the prediction accuracy and develop more robust prediction models by adopting artificial intelligence methods, and clinical validation of the data analytics.

ACKNOWLEDGMENT

The research leading to these results have received funding from the European Commission funded program under H2020 Grant Agreement No: 644906 (AEGLE), <http://www.aegle-uhealth.eu>

REFERENCES

- [1] D. Soudris, S. Xydis, C. Baloukas, A. Hadzidimitriou, I. Chouvarda, K. Stamatopoulos, N. Maglaveras, N. Philip, J. Chang, A. Raptopoulos, D. Manset et al., "AEGLE: A big bio-data analytics framework for integrated health-care services," in 2015 International Conference on Embedded Computer Systems: Architectures, Modeling, and Simulation (SAMOS). IEEE, 2015, pp. 246–253.
- [2] N. Holman, B. Young, and R. Gadsby, "Current prevalence of type 1 and type 2 diabetes in adults and children in the uk," *Diabetic Medicine*, vol. 32, no. 9, pp. 1119–1120, 2015.
- [3] "Number of people with diabetes reaches 4.7 million," https://www.diabetes.org.uk/about_us/news/new-stats-People-living-with-diabetes, accessed: 2019-10-30.
- [4] C. D. Mathers and D. Loncar, "Projections of global mortality and burden of disease from 2002 to 2030," *PLoS medicine*, vol. 3, no. 11, p. e442, 2006.
- [5] A. D. Association et al., "Economic costs of diabetes in the us in 2017," *Diabetes care*, vol. 41, no. 5, pp. 917–928, 2018.
- [6] J. Rumbold, M. O'Kane, N. Philip, and B. Pierscionek, "Big data and diabetes: the applications of big data for diabetes care now and in the future," *Diabetic Medicine*, 2019.
- [7] J. Hippisley-Cox and C. Coupland, "Development and validation of risk prediction equations to estimate future risk of blindness and lower limb amputation in patients with diabetes: cohort study," *BMJ*, vol. 351, p. h5441, 2015.
- [8] I. Marzona, F. Avanzini, G. Lucisano, M. Tettamanti, M. Baviera, A. Nicolucci, M. C. Roncaglioni et al., "Are all people with diabetes and cardiovascular risk factors or microvascular complications at very high risk? findings from the risk and prevention study," *Acta diabetologica*, vol. 54, no. 2, pp. 123–131, 2017.
- [9] S. Basu, J. B. Sussman, S. A. Berkowitz, R. A. Hayward, and J. S. Yudkin, "Development and validation of risk equations for complications of type 2 diabetes (recode) using individual participant data from randomised trials," *The Lancet Diabetes & Endocrinology*, vol. 5, no. 10, pp. 788–798, 2017.
- [10] E. B. Schroeder, S. Xu, G. K. Goodrich, G. A. Nichols, P. J. O'Connor, and J. F. Steiner, "Predicting the 6-month risk of severe hypoglycemia among adults with diabetes: Development and external validation of a prediction model," *Journal of diabetes and its complications*, vol. 31, no. 7, pp. 1158–1163, 2017.
- [11] N. Barakat, A. P. Bradley, and M. N. H. Barakat, "Intelligible support vector machines for diagnosis of diabetes mellitus," *IEEE transactions on information technology in biomedicine*, vol. 14, no. 4, pp. 1114–1120, 2010.
- [12] B. Liu, Y. Li, S. Ghosh, Z. Sun, K. Ng, and J. Hu, "Complication risk profiling in diabetes care: a bayesian multi-task and feature relationship learning approach," *IEEE Transactions on Knowledge and Data Engineering*, 2019.
- [13] A. Pavate and N. Ansari, "Risk prediction of disease complications in type 2 diabetes patients using soft computing techniques," in 2015 Fifth International Conference on Advances in Computing and Communications (ICACC). IEEE, 2015, pp. 371–375.

- [14] J. Yan, X. Du, Y. Yu, and H. Xu, "Establishment of risk prediction model for retinopathy in type 2 diabetic patients," in *International Conference on Brain Informatics*. Springer, 2019, pp. 233–243.
- [15] A. Dagliati, S. Marini, L. Sacchi, G. Cogni, M. Teliti, V. Tibollo, P. De Cata, L. Chiovato, and R. Bellazzi, "Machine learning methods to predict diabetes complications," *Journal of diabetes science and technology*, vol. 12, no. 2, pp. 295–302, 2018.
- [16] K. V. Dalakleidi, K. Zarkogianni, V. G. Karamanos, A. C. Thanopoulou, and K. S. Nikita, "A hybrid genetic algorithm for the selection of the critical features for risk prediction of cardiovascular complications in type 2 diabetes patients," in *13th IEEE International Conference on BioInformatics and BioEngineering*. IEEE, 2013, pp. 1–4.
- [17] R. Gargeya and T. Leng, "Automated identification of diabetic retinopathy using deep learning," *Ophthalmology*, vol. 124, no. 7, pp. 962–969, 2017.
- [18] M.-H. Hsieh, L.-M. Sun, C.-L. Lin, M.-J. Hsieh, K. Sun, C.-Y. Hsu, A.-K. Chou, and C.-H. Kao, "Development of a prediction model for colorectal cancer among patients with type 2 diabetes mellitus using a deep neural network," *Journal of clinical medicine*, vol. 7, no. 9, p. 277, 2018.
- [19] Y. Cheng, F. Wang, P. Zhang, and J. Hu, "Risk prediction with electronic health records: A deep learning approach," in *Proceedings of the 2016 SIAM International Conference on Data Mining*. SIAM, 2016, pp. 432–440.
- [20] D. Masouros, K. Koliogeorgi, G. Zervakis, A. Kosvira, A. Chytas, S. Xydis, I. Chouvarda, and D. Soudris, "Co-design implications of costeffective on-demand acceleration for cloud healthcare analytics: The aegle approach," in *2019 Design, Automation & Test in Europe Conference & Exhibition (DATE)*. IEEE, 2019, pp. 622–625.
- [21] M. Richmond, "Population pyramids," Oregon State University, 2014.
- [22] "rCharts by Ramnath Vaidyanathan," <https://ramnathv.github.io/rCharts>, accessed: 2019-10-25.
- [23] "Highcharts," <https://www.highcharts.com/demo/bar-negative-stack>, accessed: 2019-10-25.
- [24] Y. Zheng, S. H. Ley, and F. B. Hu, "Global aetiology and epidemiology of type 2 diabetes mellitus and its complications," *Nature Reviews Endocrinology*, vol. 14, no. 2, p. 88, 2018.
- [25] S. Finnikin, R. Ryan, and T. Marshall, "Statin initiations and QRISK2 scoring in UK general practice: a THIN database study," *Br J Gen Pract*, vol. 67, no. 665, pp. e881–e887, 2017.
- [26] L. Abarca-Gómez, Z. A. Abdeen, Z. A. Hamid, N. M. Abu-Rmeileh, B. Acosta-Cazares, C. Acuin, R. J. Adams, W. Aekplakorn, K. Afsana, C. A. Aguilar-Salinas et al., "Worldwide trends in body-mass index, underweight, overweight, and obesity from 1975 to 2016: a pooled analysis of 2416 population-based measurement studies in 128.9 million children, adolescents, and adults," *The Lancet*, vol. 390, no. 10113, pp. 2627–2642, 2017.
- [27] D. B. Rubin and N. Schenker, "Multiple imputation in health-care databases: An overview and some applications," *Statistics in medicine*, vol. 10, no. 4, pp. 585–598, 1991.
- [28] S. van Buuren, "Multivariate imputation by chained equations."
- [29] A. B. Pedersen, E. M. Mikkelsen, D. Cronin-Fenton, N. R. Kristensen, T. M. Pham, L. Pedersen, and I. Petersen, "Missing data and multiple imputation in clinical epidemiological research," *Clinical epidemiology*, vol. 9, p. 157, 2017.
- [30] P. C. Austin, "Generating survival times to simulate cox proportional hazards models with time-varying covariates," *Statistics in medicine*, vol. 31, no. 29, pp. 3946–3958, 2012.
- [31] H. Uno, T. Cai, L. Tian, and L.-J. Wei, "Evaluating prediction rules for t-year survivors with censored regression models," *Journal of the American Statistical Association*, vol. 102, no. 478, pp. 527–537, 2007.
- [32] "Package survAUC," <https://cran.r-project.org/web/packages/index.html>, accessed: 2019-10-20.
- [33] T. Bejan-Angoulvant, C. Cornu, P. Archambault, B. Tudrej, P. Audier, Y. Brabant, F. Gueyffier, and R. Boussageon, "Is HbA1c a valid surrogate for macrovascular and microvascular complications in type 2 diabetes?" *Diabetes & metabolism*, vol. 41, no. 3, pp. 195–201, 2015.
- [34] E. Ahlqvist, P. Storm, A. Käräjämäki, M. Martinell, M. Dorkhan, A. Carlsson, P. Vikman, R. B. Prasad, D. M. Aly, P. Almgren et al., "Novel subgroups of adult-onset diabetes and their association with outcomes: a data-driven cluster analysis of six variables," *The Lancet Diabetes & Endocrinology*, vol. 6, no. 5, pp. 361–369, 2018.
- [35] H.-Y. Tsao, P.-Y. Chan, and E. C.-Y. Su, "Predicting diabetic retinopathy and identifying interpretable biomedical features using machine learning algorithms," *BMC bioinformatics*, vol. 19, no. 9, p. 195, 2018.



NADA Y. PHILIP is an Associate Professor in the field of Mobile health (mHealth) at Kingston University London. She is the founder of the research group Digital Media for Health in 2012. She obtained her PhD in mHealth with the thesis title 'Medical Quality of Service for Optimised Ultrasound Streaming in Wireless Robotic Teleultrasonography System' from Kingston University, UK in 2008. Her research interests are mainly in the advancement of Data and Multimedia Communication, Networking and Information Technology for healthcare applications. She is the PI and Co-PI of many national and international mHealth projects in the areas of personalized health for Diabetes, Cancer and COPD conditions, 5G health, wearables and cloud computing, IoT, AI and Big data analytics for health, social robotics for health, end to end QoS and QoE in medical video streaming. She is the author and co-author of more than 80 journals, peer reviewed conferences and book chapters. She is a member of the editorials and the review panels for many journals including IEEE-IoT, JSAC and WCMC. She is the editor of the IEEE e-health TC newsletter. She is a reviewer on both the MRC and the NIHR research bodies. She is a fellow of the Higher Education Academy, Senior Member of the IEEE, IEEE communication society and the IEEE Engineering in Medicine and Biology.



MANZOOR RAZAAK received the B.E. degree in electronics and communication engineering from the Sambhram Institute of Technology, Bangalore, India, in 2010, and the M.Sc. degree in embedded systems from Kingston University, London, U.K, in 2011. From the same university he also received the Ph.D degree in computer science, in 2016. After PhD graduation, he worked as a post-doctoral research associate at Kingston University London for four years before moving on to head the R&D team at a medical imaging company. His research interests include computer vision, deep learning, medical image processing, data analysis, and image processing.



JOHN CHANG graduated from Bristol University in the UK with a BSc in medical microbiology as well as MB ChB in medicine. After graduation he started his NHS training in the South West England region before moving north to Leicestershire, then Sheffield, before finally moving back to London. During his registrar rotation in London, he took a year out in Brisbane Mater Mother's hospital, working under Dr Gray, and started his research interest: use of Doppler flow studies in preterm infants, as well as use of nasal CPAP to support newborn. After his return to the UK, he secured a Consultant post at Croydon in 1993, as lead for the sub regional neonatal Intensive Care unit, and Honorary Senior Lecturer post at St George's medical school. He continued to be involved in both Ethics and research, becoming Chair of the Local Research Ethics Committee before their replacement by MRECs and the IRAS system, as well as being an R&D committee member before being appointed the R&D director and head of the R&D office at Croydon since 2001. During the period in Croydon, he has been involved in numerous multicentre studies as well as own locally generated research within the unit, in addition to overseeing the research developments within Croydon University Hospital. As part of the R&D role, he has been involved in numerous collaborations with researchers to help secure grants, working with both NIHR, Innovate UK as well as the EU horizon 2020 programme. Highlights of projects linked to Croydon include Welcome wearables as part of EU FY7 study; AEGLE as part of Horizon 2020, Optimal as part of Innovate UK; FIT trial as part of London Cancer Alliance, and OSIRIS trial as part of Health Foundation and the Royal College of Obstetrics and Gynaecology. The Trust has also secured some commercial research projects looking at use of copper in terms of infection control. He also sits on the Clinical Research Partnership Board, and the South London Clinical Research Network Board. As a result of the various projects, he has published widely in numerous journals, as well as presented at regional, national and international meetings. He remains an active researcher and collaborator within the NHS at Croydon.



SUCHETHA M Senior Member, IEEE received Ph.D in Biomedical with the thesis title 'Empirical Mode Decomposition based

denoising and classification techniques applied to Electrocardiogram signals'. She is currently Deputy Director, Centre for Healthcare advancements, Innovation and Research in VIT University. Her areas of interest are wearable devices, Signal and image processing techniques in biomedical, AI and Big data analytics for healthcare, developing non-invasive devices. She is the Principal Investigator for a funded project by ISRO and other funded projects. She has published more than 70 papers in journals, authored 4 books and filed several patents. She is a life member of ISTE, Senior member of IEEE Engineering in Medicine and Biology Society (EMBS) and Faculty advisor of IEEE Robotics and Automation society.



MAURICE O'KANE graduated in medicine at the University of Edinburgh and undertook postgraduate training in Scotland, N. Ireland and France, firstly in internal medicine and then in chemical pathology. He has been a consultant chemical pathologist in the Western Health and Social Care Trust, in N. Ireland for over 25 years where he is head of the clinical biochemistry laboratory service, including point-of-care

testing and with

clinical interests in lipid disorders and diabetes mellitus. Major research interests include point of care testing, patient self-management in chronic disease and the evaluation and the adoption / implementation of diagnostic tests. Maurice is currently Joint Editor-in-Chief of the Annals of Clinical Biochemistry and Director of the N. Ireland Clinical Research Network. He has been visiting professor in biomedical sciences at Ulster University for many years. Previous professional responsibilities have included Director of Clinical Practice at the Association for Clinical Biochemistry and Laboratory Medicine (2014-2019) and Director of Research at the Western Health and Social Care Trust (2009-2016). He has over 90 publications and has written 3 book chapters.



BARBARA K PIERSCIONEK has expertise in optics, biochemistry, biomechanics, nanotechnology, cell biology and how these apply to the study of eye development, ageing and disease as well as to new technologies for sight improvement.

She is also qualified in law and has research interests in ethico-legal aspects of Big Data and emerging technologies. She graduated from Melbourne University in Australia with a PhD on the protein chemistry and optics of the eye lens and was awarded a prestigious NHMRC research fellow (MRC equivalent) shortly after graduating to start an independent research program on the optics of the eye. She led the Vision Research group in Biomedical Science at Ulster University and worked as the Associate Dean (Research and Enterprise) in the Faculty of Science, Engineering and Computing at Kingston University. Subsequently, she held dual roles at Nottingham Trent University: Associate Dean Research in the School of Science and Technology and pan-Institutional Head of

Health and Wellbeing. More recently she led research in two Schools as Associate Dean Research in the School of Life Science and Education and in the School of Health and Social Care at Staffordshire University before taking up the post of Deputy Dean (Research and Innovation) in the Faculty of Health, Education, Medicine and Social Care at Anglia Ruskin University. She has continued as an active researcher in the areas of optics and biomechanics of the eye, ageing of the eye and eye disease, bioinformatics, sports and binocular vision, nanotechnological applications to the eye and ethico-legal aspects of Big Data. She has received support for her research with funding from Research Councils (EPSRC, BBSRC), EU, Fight for Sight, Royal Society and industry (Essilor International and Zeiss Meditec) as well as being awarded beam time grants for work in Japan at SPring-8 the world's largest synchrotron. She has published over 150 peer review papers, 4 book chapters and a book on law and ethics for the eye care practitioner.