Computers in Biology and Medicine 75 (2016) 90–97

Contents lists available at ScienceDirect

Computers in Biology and Medicine

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

Data analytics identify glycated haemoglobin co-markers for type 2 diabetes mellitus diagnosis

Herbert F. Jelinek [a], Andrew Stranieri [b,*], Andrew Yatsko [b], Sitalakshmi Venkatraman [b,c]

[a] School of Community Health and Centre for Research in Complex Systems, Charles Sturt University, PO Box 789, Albury, NSW 2640, Australia
[b] Centre for Informatics and Applied Optimisation, Federation University, PO Box 663, University Drive, Mt Helen, Victoria 3350, Australia
[c] Department of Higher Education – Business (IT), Melbourne Polytechnic, 77-91 St Georges Rd, Preston, Victoria 3072, Australia

ARTICLE INFO

Article history:
Received 22 November 2015
Received in revised form
10 May 2016
Accepted 12 May 2016

Keywords:
Glycated haemoglobin
Type 2 diabetes mellitus
Data analytics

ABSTRACT

Glycated haemoglobin (HbA1c) is being more commonly used as an alternative test for the identification of type 2 diabetes mellitus (T2DM) or to add to fasting blood glucose level and oral glucose tolerance test results, because it is easily obtained using point-of-care technology and represents long-term blood sugar levels. HbA1c cut-off values of 6.5% or above have been recommended for clinical use based on the presence of diabetic comorbidities from population studies. However, outcomes of large trials with a HbA1c of 6.5% as a cut-off have been inconsistent for a diagnosis of T2DM. This suggests that a HbA1c cut-off of 6.5% as a single marker may not be sensitive enough or be too simple and miss individuals at risk or with already overt, undiagnosed diabetes. In this study, data mining algorithms have been applied on a large clinical dataset to identify an optimal cut-off value for HbA1c and to identify whether additional biomarkers can be used together with HbA1c to enhance diagnostic accuracy of T2DM. T2DM classification accuracy increased if 8-hydroxy-2-deoxyguanosine (8-OhdG), an oxidative stress marker, was included in the algorithm from 78.71% for HbA1c at 6.5% to 86.64%. A similar result was obtained when interleukin-6 (IL-6) was included (accuracy=85.63%) but with a lower optimal HbA1c range between 5.73 and 6.22%. The application of data analytics to medical records from the Diabetes Screening programme demonstrates that data analytics, combined with large clinical datasets can be used to identify clinically appropriate cut-off values and identify novel biomarkers that when included improve the accuracy of T2DM diagnosis even when HbA1c levels are below or equal to the current cut-off of 6.5%.

1. Introduction

Type 2 diabetes mellitus (T2DM) is anticipated to rise to 366 million people by 2030 [1]. T2DM progression can be reduced if recognised early and appropriately treated. Emerging evidence from clinical trials has demonstrated the effectiveness of targeted population screening. In Europe, a randomized controlled trial associated a small, non-significant reduction in the incidence of cardiovascular events and death with early screening and intensive multifactorial intervention for diabetes [2].

Effective screening requires accurate models for diagnosing T2DM, which may include inflammation and oxidative stress factors, known to increase risk of T2DM. Currently a clinical diagnosis of diabetes is based on a set of guidelines that specify levels of impaired fasting glucose (IFG) and impaired glucose tolerance (IGT); however diagnosing T2DM with these factors alone is problematic [3]. Glycated haemoglobin (HbA1c) has been recommended as an alternative or additional clinical test for diagnosis of T2DM [2,4,5]. An Australian screening study has highlighted the effectiveness of including HbA1c as an indicator for the need to follow-up with an Oral Glucose Tolerance Test (OGTT) in addition to IFG in determining T2DM [6]. However HbA1c has been shown to miss a substantial section of the population with overt T2DM [7].

HbA1c reflects average blood glucose levels (BGL) over the preceding 6–12 weeks reflecting the average life span of the red blood cells. HbA1c is a convenient point-of-care test as fasting is not required. It has several advantages to FPG and OGTT results providing a long-term indicator for blood glucose level (BGL) if erythrocyte function is normal [8]. In addition, HbA1c is more stable and preserves better from the time of collection to the time of assay, and also has less specific storage requirements, compared with glucose [5]. Threshold levels of HbA1c for T2DM diagnosis vary across countries but 6.5% (48 mmol/L) is the recommended cut-off by the American Diabetes Association (ADA) [8]. Although HbA1c is convenient, its effectiveness in identifying T2DM has been questioned. In the United States, approximately 25% of people have undiagnosed diabetes as BGL or HbA1c have not reached the clinical cut-off value and in certain individuals there is an incomplete correlation between HbA1c and average BGL within a

[*] Corresponding author.

http://dx.doi.org/10.1016/j.compbiomed.2016.05.005
0010-4825/© 2016 Elsevier Ltd. All rights reserved.

wide spectrum of low-risk to high-risk clinical presentations [8].

Sensitivity and specificity of T2DM diagnosis may plausibly be improved if levels of other biomarkers, demographic and lifestyle factors are coupled with low HbA1c values. Inflammatory and oxidative stress markers, which are two pathophysiological processes known to be involved in T2DM progression may provide such information [9,10]. Clinical and demographic features such as gender, age, race/ethnicity, anaemia/hemoglobinopathies, body mass index (BMI) and sedentary lifestyles have been found to be risk factors for diabetes [11,12]. Also, there are conditions which are known to accompany T2DM such as cardiac autonomic neuropathy (CAN) and diabetic peripheral neuropathy (DPN). CAN is commonly diagnosed using the Ewing scoring system [13] assessing heart rate (HR) and blood pressure (BP), and a similar approach can be exercised with DPN, particularly by assessing the ankle and knee reflexes and sensory function in the feet [14]. Further, heart rate variability (HRV) measures extracted from electrocardiogram (ECG) are being actively researched as an alternative to the Ewing battery for identification of CAN [15,16]. Biochemical markers such as 8-hydroxy-2-deoxyguanosine (8-OHdG) and homocysteine (Hcy) have been reported to aid in the assessment of an individual's risk of prediabetes and developing overt diabetes [17]. Total body fat has also been suggested as important in assessing the personalised diabetes risk levels [18].

Challenges that emerge when attempting to couple HbA1c with other indicators to enhance sensitivity and specificity include the need to discover which factors may lead to improvements, and what threshold levels on those factors lead to improvements. Data analytics approaches for diabetes care and treatment are surveyed in [19]. To date, data mining techniques have not been applied for the discovery of cut-off thresholds for emerging biomarkers for T2DM such as oxidative stress and inflammation markers in conjunction with HbA1c.

The discovery of ideal cut-off values for HbA1c under different clinical circumstances that may impact on the HbA1c value are well suited to discovery by data mining techniques. Diabetes has a strict BGL cut-off value for diagnosis but a potentially long preclinical period where BGL levels have not exceeded the critical threshold but complications of diabetes can already start to manifest during this period [7]. The aim of the current study was to discover possible biomarkers and respective cut-off thresholds that, alongside HbA1c, could predict T2DM. The use of data analytics in this way is envisaged as a precursor to controlled population based studies to validate the markers and thresholds. The approach involved the application of predictive analytics to a large rural clinical dataset where participants had a wide range of indicators including HbA1c measured as part of a cardiovascular and diabetes screening unit involvement [14].

2. Materials and methods

2.1. Clinical Dataset

The application of data mining techniques for the discovery of markers and thresholds to be used in conjunction with HbA1c requires a dataset that contains data from participants on a large number of biomarkers, demographics and other factors to identify the combinations of markers and cut-off thresholds that best predict T2DM. In addition, a representation for depicting the combinations of markers and thresholds that is readily understood by clinicians and researchers becomes necessary if combinations are numerous.

The dataset used in this study was derived from the Diabetes Health screening (DiabHealth) conducted at a regional Australian university [14]. The study was approved by the Charles Sturt University Human Ethics Committee. All participants provided informed, written consent. The DiabHealth community screening concentrates on diabetes, cardiovascular disease and hypertension as a triad of diseases, which are specifically identified in the dataset. The screening clinic has been collecting data over ten years and includes information such as demographics, socio-economic variables, and clinical variables such as BGL, HbA1c, cholesterol profile, inflammatory and oxidative stress markers, as well as an extensive medical history, BMI, peripheral vascular function, and ECG derived variables. Data on 300 attributes from 2860 attendances by 840 patients have been collected in recent years. Demographic, clinical and biological samples for subsequent pathology testing were collected at the diabetes health screening clinic (DiabHealth).

The dataset has been used in several data mining applications for identification of novel data mining algorithms and diabetes disease classification [20–23]. Patients with Type 1 diabetes mellitus (T1DM) were excluded in the current study. Participants were recorded to have T2DM if they reported having had a clinical diagnosis prior to the screening or taking glucose-lowering medication. Participants who did not report a T2DM but had an HbA1c score above 6.5% or elevated fasting BGL (> 7 mmol/L) were also recorded as having T2DM.

2.2. Novel data analytics

The final diabetes dataset assembled for this study was a subset of the DiabHealth data and contained 99 attributes selected from the 300 available if there was some indication that the attribute may plausibly enhance T2DM accuracy when coupled with HbA1c. T2DM was the class attribute to be classified. Patients diagnosed with T2DM represented 26% of the records (labelled Class 1), while non-diabetic patients (Class 0) represented the control class. Missing values (MVs) were imputed in order to apply data mining algorithms to complete data for reasons outlined by Jelinek et al. [22]. In this work, the modified General Location Model (GLM) approach was used to impute MVs following Baghirov et al. [23]. The novel data analytics technique developed for the current research is described in the sequence of steps described below. Computer programs for each step below and the pre-processing required were custom written in the C language. Information gain was calculated using the Shannon information gain equation embedded in standard decision tree algorithms including ID3 and C5.0.

2.2.1. Step 1. Generate an initial decision tree

Using T2DM as the class feature (Class 1), a conventional decision tree is first generated using an information gain (IG) measure. Leaf nodes are the T2DM outcomes. A decision tree was preferred over ‘black box’ classifiers such as neural networks because influential variables could readily be recognised in a decision tree. Other classifiers such as support vector machines offer similar accuracies and computational efficiencies but this was not important for this study. As Fernandez-Delgado et al. [24] note, differences in accuracies between classifiers is not statistically significant when tested over many datasets.

A conventional decision tree algorithm is not guaranteed to position HbA1c at the root of the tree as other features may provide greater information gain. However, HbA1c needs to be placed at the root of the tree in this study to allow HbA1c to be the main marker and identify additional biomarkers. Generating a decision tree with HbA1c as root was achieved by removing BGL in the dataset as this factor is the main clinical decision variable and hence distinguished T2DM from no T2DM better than HbA1c. The decision tree with HbA1c at the root also identified a cut-off threshold that maximised information gain.