ELSEVIER

Archives of Medical Research ■ (2018) ■

Archives of Medical Research

OPINION

Data Mining and Endocrine Diseases: A New Way to Classify?

Juan Salazar,[a,*] Cristobal Espinoza,[b,*] Andres Mindiola,[c,*] and Valmore Bermudez[d,*]

[a]*Centro de Investigaciones Endocrino Metabólicas Dr. Félix Gómez, Escuela de Medicina, Universidad de Zulia, Maracaibo, Venezuela*
[b]*Hospital General Provincial de Latacunga, Ministerio de Salud Pública, Cotopaxi, Ecuador*
[c]*Department of Pathology and Laboratory Medicine, Dartmouth-Hitchcok Medical Center, Lebanon, NH, USA*
[d]*Universidad Simón Bolívar, Facultad de Ciencias de la Salud, Barranquilla, Colombia*

Received for publication July 14, 2018; accepted August 8, 2018 (ARCMED_2018_212).

Data mining consists of using large database analysis to detect patterns, relationships and models in order to describe (or even predict) the appearance of a future event; to accomplish this, it uses classification methods, rules of association, regression patterns, link and cluster analyses. Recently this approach has been used to propose a new diabetes mellitus classification, using information analysis techniques through which the selection bias minimally influences categorization, this new focus that includes data mining previously implemented to predict, identify biomarkers, complications, therapies, health policies, genetic and environmental effects of this disease; it could be generalized in the field of endocrinology, in the classification of other endocrine diseases. © 2018 IMSS. Published by Elsevier Inc.

Key Words: Data mining, Classification, Endocrine disease, Diabetes mellitus, Information analysis.

Currently computer systems have a great part on our daily living, their appliance and automation in health has become a growing tendency in developed countries; this may influence patient diagnostic and therapeutic decisions or even work as organization method in great scale among health systems (1).

Data mining consists of using large database analysis to detect patterns, relationships and models in order to describe (or even predict) the appearance of a future event; to accomplish this, it uses classification methods, rules of association, regression patterns, link and cluster analyses (2). In healthcare each action or decision represents data and data mining constitutes a promising technique for information analysis.

Multiple reports have shown that not only specific disease prediction could be managed through machine learning; but also physical conditions such as personal wellness. For example, Semerdjian and Frank reported a high capacity to predict Type 2 Diabetes Mellitus (T_2DM) with a data analysis in the National Health and Nutrition Survey (NHANES), they employed a model starting from a group of algorithms and 16 classifying variables (3). Argawal et al., employed a survey by the Center for Disease Control and Prevention to make predictive models of health status in a wide group of subjects and determine their level of wellness (4).

To this date, numerous chronic diseases considered as the main causes of morbi-mortality in the adult population, have been defined or classified based on criteria from more than 50 years old data. Most of this data has an arbitrary origin or may be just not adjusted to the population ethnic context. In this sense, Diabetes Mellitus (DM) is an endocrine disease with their first classification indices come from the 1930s. In this age, Himsworth starts to measure the insulin effectiveness (5), since then research around DM physiopathology has highly progressed. However, traditional classification is still being used in the clinical setting, probably because its simplicity for clinical practice.

Recently, Emma Ahlqvist et al. in *The Lancet Diabetes & Endocrinology,* reported a disease subclassification in 8980 patients based specially on T_2DM heterogeneity. This patients had a new diagnosis from the Swedish All New Diabetics in Scania cohort (6). In order to do this, they conducted a cluster analysis (k-means and hierarchical

*These authors contributed equally to this work.

Address reprint requests to: Juan Salazar, Endocrine-Metabolic Research Center, Dr. Félix Gómez, 20th Avenue, Maracaibo 4004, Venezuela, Bolivarian Republic of; Phone/FAX: (+58) (261) 7597279; E-mail: juanjsv18@hotmail.com

0188-4409/$ - see front matter.
https://doi.org/10.1016/j.arcmed.2018.08.005

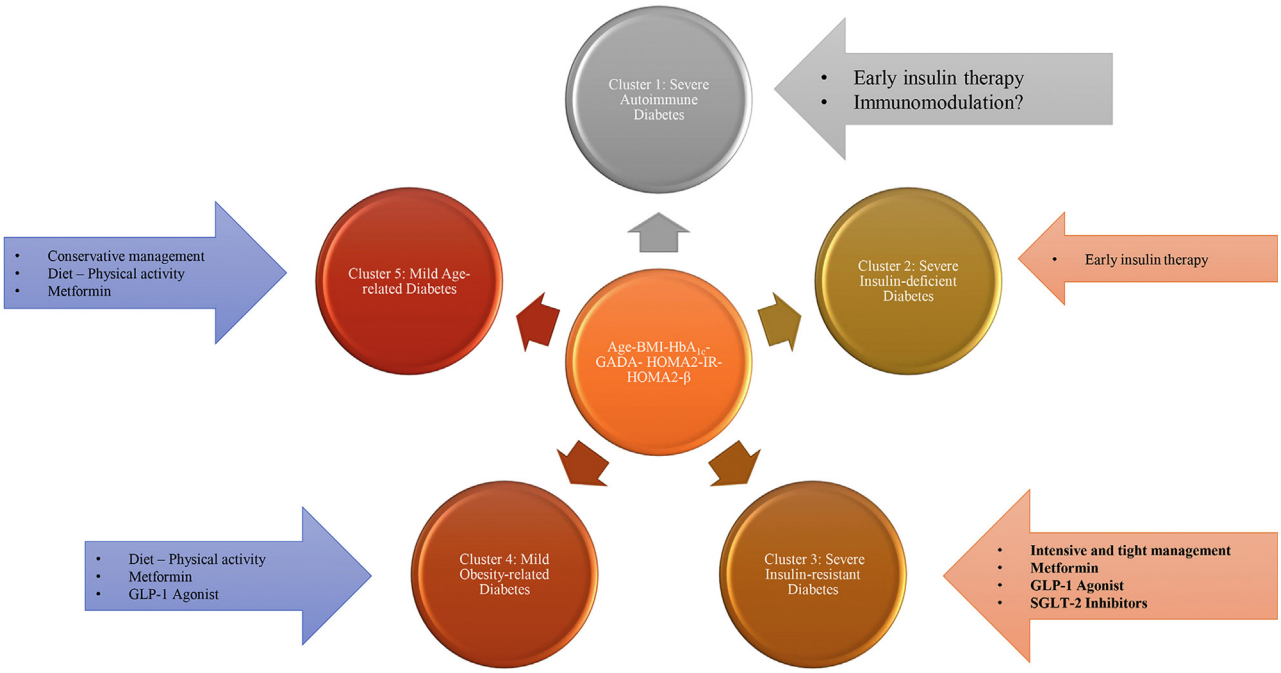

Figure 1. Subclassification of diabetic subjects according to the All New Diabetics in Scania cohort and potential therapeutic approaches. Based on subclassificationgenerated from the cluster analysis of All New Diabetics in Scania cohort, the potential therapeutic strategies in each groups are considered, considering the limited current pharmacological arsenal with sufficient level of evidence for patients with more complex management. BMI: Body Mass Index; GADA: Glutamate decarboxylase antibodies; HbA1c: Glycated haemoglobin A1C.

clustering) using six pre-established parameters: age at diagnosis, Body Mass Index, HbA1c, β-cell function estimates by homeostatic model assessment 2, insulin resistance, and glutamate decarboxylase antibodies; the latter are usually difficult to measure in daily clinical practice. From this analysis, they obtained five diabetic patient clusters with different features from clinical, therapeutic and risk of complications point of view (Figure 1). The fact that this results has been obtained in other Swedish cohorts and prospectively evaluated in terms of time to medication, time to reaching the treatment goal, risk of diabetic complications, and genetic association results interesting.

Considering the interesting Swedish findings, many questions and proposals have originated. For example, do we have the therapeutic arsenal to cover the physiopathology heterogeneity coming with these new patterns? Until now, it seems that early insulin therapy would be the first option for those with insulin deficit secretion since the beginning of the disease. A more conservative treatment based on an intense non pharmacologic intervention would be better for patients classified in clusters four and five, which may even include drugs that modulate energy consumption in obese patients. Patients with insulin resistance and potential complication to specific organs, have worse prognosis and require an intervention that must be opportune, intensive, tight controlled and with adequate adherence from the patient (Figure 1).

In this sense, the effect of other acquired risk factors on the classification and the specific observed complications, the inclusion of genetic variables in multivariable model analysis, and its reproducibility in diverse non Caucasian ethnic groups (considering population variations in multiple metabolic variables), represents few of many aspects to consider in future longitudinal research on large populations. This new focus that includes data mining is not only limited to predict DM, but also to identify biomarkers, complications, therapies, health policies, genetic and environmental effects (7); and now its own classification. Data mining approaches has not only been explored for DM but it has also been applied to other endocrine diseases such as obesity (8), thyroid diseases (9), and polycystic ovary syndrome (10).

References

1. Kumar R, Shaikh BT, Chandio AK, et al. Role of Health Management Information System in disease reporting at a rural district of Sindh. Pak J Public Health 2012;2:10–12.
2. ŢĂranu I. Data mining in healthcare: decision making and precision. Database Systems Journal 2015;VI:33–40.
3. Semerdjian J, Frank S. An Ensemble Classifier for Predicting the Onset of Type II Diabetes. arXiv:1708.07480. Available: https://arxiv.org/pdf/1708.07480.pdf. Accessed June 15, 2018.
4. Agarwal A, Baechle C, Behara RS, Rao V. Multi-method approach to wellness predictive modeling. J Big Data 2016;3:15.

5. Himsworth HP. The syndrome of diabetes mellitus and its causes. Lancet 1949;1:465–473.
6. Ahlqvist E, Storm P, Käräjämäki A, et al. Novel subgroups of adult-onset diabetes and their association with outcomes: a data-driven cluster analysis of six variables. Lancet Diabetes Endocrinol 2018;6:361–369.
7. Kavakiotis I, Tsave O, Salifoglou A, et al. Machine Learning and Data Mining Methods in Diabetes Research. Comput Struct Biotechnol J 2017;15:104–116.
8. Bermúdez V, Rojas J, Salazar J, et al. Sensitivity and Specificity Improvement in Abdominal Obesity Diagnosis Using Cluster Analysis during Waist Circumference Cut-Off Point Selection. J Diabetes Res 2015;2015:750265.
9. Jajroudi M, Baniasadi T, Kamkar L, Arbabi F, Sanei M, Ahmadzade M. Prediction of survival in thyroid cancer using data mining technique. Technol Cancer Res Treat 2014;13:353–359.
10. Dewailly D, Alebić MŠ, Duhamel A, Stojanović N. Using cluster analysis to identify a homogeneous subpopulation of women with polycystic ovarian morphology in a population of non-hyperandrogenic women with regular menstrual cycles. Hum Reprod 2014;29:2536–2543.