Journal of Biomedical Informatics 44 (2011) 319–325

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

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

Data mining methods for classification of Medium-Chain Acyl-CoA dehydrogenase deficiency (MCADD) using non-derivatized tandem MS neonatal screening data

Tim Van den Bulcke [a,*], Paul Vanden Broucke [a], Viviane Van Hoof [b,c,d], Kristien Wouters [a], Seppe Vanden Broucke [a], Geert Smits [a], Elke Smits [a], Sam Proesmans [d], Toon Van Genechten [d], François Eyskens [b,d,e]

[a] *i-ICT, University Hospital Antwerp, Wilrijkstraat 10, 2650 Edegem, Belgium*
[b] *Provinciaal Centrum voor de Opsporing van Metabole Aandoeningen (PCMA), Antwerpen, Belgium*
[c] *Dept. of Clinical Chemistry, University Hospital Antwerp, Edegem, Belgium*
[d] *Faculty of Medicine, University Antwerp, Antwerpen, Belgium*
[e] *Dept. of Paediatrics/Metabolic Diseases, University Hospital Antwerp, Edegem, Belgium*

ARTICLE INFO

Article history:
Received 15 July 2010
Available online 15 December 2010

Keywords:
Data mining
Rare diseases
MCADD
Medium-Chain Acyl-CoA dehydrogenase
Logistic regression

ABSTRACT

Newborn screening programs for severe metabolic disorders using tandem mass spectrometry are widely used. Medium-Chain Acyl-CoA dehydrogenase deficiency (MCADD) is the most prevalent mitochondrial fatty acid oxidation defect (1:15,000 newborns) and it has been proven that early detection of this metabolic disease decreases mortality and improves the outcome. In previous studies, data mining methods on derivatized tandem MS datasets have shown high classification accuracies. However, no machine learning methods currently have been applied to datasets based on non-derivatized screening methods.

A dataset with 44,159 blood samples was collected using a non-derivatized screening method as part of a systematic newborn screening by the PCMA screening center (Belgium). Twelve MCADD cases were present in this partially MCADD-enriched dataset. We extended three data mining methods, namely C4.5 decision trees, logistic regression and ridge logistic regression, with a parameter and threshold optimization method and evaluated their applicability as a diagnostic support tool. Within a stratified cross-validation setting, a grid search was performed for each model for a wide range of model parameters, included variables and classification thresholds.

The best performing model used ridge logistic regression and achieved a sensitivity of 100%, a specificity of 99.987% and a positive predictive value of 32% (recalibrated for a real population), obtained in a stratified cross-validation setting. These results were further validated on an independent test set. Using a method that combines ridge logistic regression with variable selection and threshold optimization, a significantly improved performance was achieved compared to the current state-of-the-art for derivatized data, while retaining more interpretability and requiring less variables. The results indicate the potential value of data mining methods as a diagnostic support tool.

1. Introduction

The early diagnosis of rare diseases constitutes a great challenge in current medicine. Currently rare diseases are often diagnosed too late, resulting in a decline in life expectancy and quality of life, and an increase in healthcare costs. The large variety of rare diseases makes that – although each disease affects only a small number of people – it still affects a large population. As such, the total number of patients suffering from a rare disease in Europe is around 30 million [1]. Early detection and diagnosis of metabolic disorders and of rare diseases in general, are of crucial importance for the further outcome of the patient. As such, statistical and machine learning methods could be of great value as a diagnostic support tool for doctors and medical personnel.

1.1. MCADD

This study focuses on MCADD (Medium-Chain Acyl-CoA dehydrogenase deficiency), the most frequent metabolic disorder of mitochondrial fatty acid oxidation [2]. There are four categories of fatty acids, differentiated by their carbon chain length: short

* Corresponding author.
E-mail address: tim.van.den.bulcke@uza.be (T. Van den Bulcke).

1532-0464/$ - see front matter © 2010 Elsevier Inc. All rights reserved.
doi:10.1016/j.jbi.2010.12.001

(C2–4), medium (C4–12), long (C12–20) and very long (>C20). The enzymes responsible for their β-oxidation are respectively SCAD, MCAD, LCAD and VLCAD (small, medium, long and very long chain acyl-Co-A dehydrogenase).

Our body preferentially metabolizes carbohydrates, but only a limited stock of carbohydrates is available and after a fasting period of 9–10 h (e.g. a normal night of sleep), the body switches to energy production from fatty acids. Defects of mitochondrial fatty acid β-oxidation therefore lead to a disturbed or inhibited energy production of fatty acids. Inherited defects fall into three groups: (a) those associated with the carnitine mediated transport into the mitochondria; (b) those of the matrix enzymes (such as MCAD); and (c) those affecting the activity of membrane-bound enzymes of long-chain fatty acid oxidation.

In case of MCADD, the production of the MCAD enzyme is absent or reduced. As such, the β-oxidation of the fatty acids C4 and higher fails and they can subsequently not be used as an energy source. The symptomatic MCADD-patient shows a clinical picture strongly resembling Reye's syndrome with hepatomegaly and stupor associated with hypoketonemia, hypoglycemia, hypocarnitinemia, increased transaminsase and mild hyperammonemia [3]. Lipids that cannot be used precipitate in liver, heart, kidneys. These patients present themselves over the years with hepatomegaly or hepatic steatosis, cardiomyopathy, encephalopathy and decreased muscle tone. 10–20% of the patients develop rhabdomyolysis in the first three years of life, even when adequately treated [4].

The early diagnosis of MCADD – and metabolic diseases in general – is crucial for the further outcome and prognosis of the patient. If the diagnosis is made early, the quality of life can be substantially improved. With supplementation of acylcarnitine and a diet high in carbohydrates and low in fats and fasting periods not longer than 6 h, the prognosis for the MCADD patient is very favorable [4]. Through early diagnosis of MCADD, the risk of death during derailment reduces to zero and the neurological rest lesions (epilepsy, paralysis, behavioral disorders, developmental disorders) after decompensation are halved [4]. There is thus an important role for preventive medicine where a metabolic disease is transformed into a metabolic disorder by means of simple measures (prevention of fasting and rapid care of sober states) that prevent the development of the disease. These children should be followed in the first 5–7 years of life to avoid decompensation. This can be done by the general practitioner and does not require a specialized center.

1.2. MCADD screening

MCADD in infants can be detected via a blood sample which is taken within a few days after birth using a heel prick test. The heel prick is performed systematically for all newborns in many developed countries (e.g. Denmark, The Netherlands, Germany, Belgium and Luxembourg). The blood sample is subsequently analyzed using tandem mass spectrometry. Depending on the screening center, a derivatized [5] or non-derivatized [6] screening method is used. A sample spectrum is shown in Fig. 1 for both a normal and an MCAD-deficient person.

An increase of the specific acylcarnitine values above established (device-specific) cut-off points usually results in a second blood analysis, carried out when the child is 8 weeks old. This second analysis includes the determination of acylcarnitine values – with tandem mass spectrometry – and the fatty acid profile in plasma. Then the organic acids and glycines in urine are determined [7]. If this second test shows no normalization of the acylcarnitine values, there is need for further review by enzymatic studies and/or DNA analysis.

Many screening centers currently use a derivatized screening method [5]. However, the PCMA as well as some other screening centers in Europe, have switched to using a non-derivatized screening method [6]. The key difference is that the derivatizaton step which requires heating of the dried analytes with dry, acidified (3 N) butanol, is no longer needed. The non-derivatized method requires less processing steps, leads to faster extraction times and has a lower cost for reagents [8].

While both methods show strong correlation among the different measured analytes, it was reported that several analytes showed consistent bias (C0, C2, C10, C16, Gly and Arg) for the non-derivatized method compared to the derivatized method [9]. For four instances this bias was due to higher recovery (C2, C10, C16 and Arg) and for the two others (C0 and Gly) this was due to a lower recovery. This bias can potentially affect the performance of data mining algorithms and may also lead to slightly different models or model parameters compared to models for derivatized data.

1.3. Data mining methods for MCADD classification

Several statistical techniques have been published to establish cutoff values on acylcarnitine values for MCADD classification [10–13]. A comparison of different data mining algorithms for classification of MCADD and other metabolic disorders on derivatized tandem MS neonatal data was done by Baumgartner et al. [14,15]. A feature selection approach for MCADD classification by Ho et al. [16] can be considered as the current state-of-the art data mining method with respect to performance. They reported sensitivity and specificity values of 100% and 99.901% respectively on a dataset of derivatized tandem MS neonatal data in Heidelberg.

This study is the first application of machine learning techniques on *non*-derivatized neonatal screening data. We applied C4.5 decision trees, logistic regression and ridge logistic regression using a grid search approach to optimize model parameter settings, included variables and classification thresholds. Our results using ridge logistic regression show a significantly better performance compared to the current state-of-the-art method for derivatized MS data [16] while our method requires less variable measurements and retains more interpretability.

2. Materials and methods

2.1. Dataset

An anonymized dataset of 44,159 blood samples was collected and analyzed using a non-derivatized tandem MS screening method [6] containing 12 MCADD cases. The dataset consists of two separate parts, each measured using a different screening system.

The *first part* is used as *training data* for learning the classification models. It was obtained as part of a systematic screening for newborns by the PCMA screening center (Belgium) during the first half of 2009. It consists of 32,109 samples and was collected using the *Quattro micro* screening system. This dataset was further enriched with blood samples of all MCADD cases that occurred between 2003 and 2009 at the PCMA screening center, resulting in a total of 9 MCADD samples. These 9 MCADD cases were further confirmed with a genetic test and to our best knowledge, no unidentified MCADD cases are present in the dataset.

The *second part* of the dataset is used as an *independent test set*. It was analyzed using a different screening system (*Xevo QT MS*) and consists of 12,050 samples (collected between June 2010 and September 2010). This dataset contained no MCADD cases from the general population and has been enriched with three spiked blood samples that were provided by the Centre for Disease