A data-mining approach to improving Polycythemia Vera diagnosis

Mehmed Kantardzic[a],[*], Benjamin Djulbegovic[b], Hazem Hamdan[a]

[a]*Computer Engineering and Computer Science Department, J.B. Speed Scientific School, University of Louisville, Louisville, KY 40292, USA*
[b]*Division of Blood and Bone Marrow Transplant, H. Lee Moffitt Cancer Center and Research Institute, University of South Florida, Tampa, FL, USA*

Abstract

This paper presents a data-mining approach to the extraction of new decision rules for Polycythemia Vera (PV) diagnosis, based on a reduced and optimized set of lab parameters. Ten laboratory and other clinical findings (eight parameters from the Polycythemia Vera Study Group (PVSG) criteria + sex and hematocrit (HCT)) on 431 PV patients from the original PVSG cohort, and records on 91 patients with other myeloproliferative disorders that can be easily misdiagnosed with PV, were included in this study. Significant differences were not found in the correctness of diagnostic classification of patients using either a trained artificial neural network (98.1%) or a support vector machine (95%) versus using PVSG diagnostic criteria, which are considered as a 'gold-standard' for the diagnosis of PV. Reducing the original parameters of our dataset to only four parameters: HCT, PLAT, SPLEEN and WBC, we still have obtained good classification results. New rules for improved differential diagnosis of PV are specified based on these four parameters. These rules may be used as a complement to the standard PVSG criteria, particularly in the differential diagnosis between PV and other myeloproliferative syndromes. © 2002 Elsevier Science Ltd. All rights reserved.

Keywords: Polycythemia Vera; Feature extraction; Artificial neural networks; Support vector machines; Decision rules; *n*-Dimensional visualization

1. Introduction

Polycythemia Vera (PV) is a rare malignant blood disease characterized by an increase in the number of red-blood cells, white-blood cells and platelets in the blood of a patient. PV represents a true neoplasm of the marrow stem cells. The diagnosis of PV is usually made by applying the widely used Polycythemia Vera Study Group (PVSG) diagnostic criteria (Berlin, 1975). Although these criteria were

* Corresponding author. Tel.: +1-502-852-3703; fax: +1-502-852-4713.
E-mail address: mmkant01@athena.louisville.edu (M. Kantardzic).

Table 1
Polycythemia Vera study group criteria PVSG for diagnosis of PV

Lab tests	
Category A	Category B
A1: Total RBC volume (REDMAS), Male $\geq$ 36 ml/kg BW, Female $\geq$ 32 ml/kg BW	B1: Thrombocytosis (WBC) $> 400,000$ mm^{-3}
A2: Arterial saturation (OSAT), O2 $> 92\%$	B2: Leukocytosis $> 12,000$ mm^{-3} (PLAT) (no fever or infection)
	B3: Leukocyte alkaline phosphatase (LAP), Score > 100
A3: Splenomegaly (SPLEEN)	B4a: Serum vitamin B12 > 900 pg/ml or (B12)
	B4b: Serum UB12BC > 2200 pg/ml (UBBC)
Diagnosis of PV	A1 + A2 + A3 or A1 + A2 + any two from category B

developed for the purpose of entering patients into clinical trials, they have been widely accepted and used in clinical practice (Djulbegovic, Hadley, & Joseph, 1991). These criteria follow the classical form of 'If…Then' categorical rules and call for the diagnosis to be established if certain clinical and laboratory data are present (Berlin, 1975; Kassirer, 1989; Kassirer & Copelman, 1989). The basic rules for the diagnosis of PV are given in Table 1. In fact, the diagnosis of PV is based on 'all or none' rule—if the criteria specified in Table 1 are present, then the diagnosis is established. If the criteria are not fulfilled, PV diagnosis is not established and may require further diagnostic testing. Consequently, treatment of PV depends on the fulfillment of this all or none categorical rule.

Despite the proven validity of the PVSG criteria (Djulbegovic et al., 1991; Djulbegovic, Hozo, & Najman, 1998), empirical data show that practitioners often do not order the tests as stipulated by the PVSG criteria. For example, a recent survey of american hematologists showed that practitioners order the determination of red-blood cell (RBC) volume in 78%, arterial gas saturation in 75%, leukocyte alkaline phosphatase (LAP) score and serum vitamin B12 level in 44% cases only during work-up of patients suspected of having PV (Streiff & Spivak, 1999).

One of the suspected reasons for this poor compliance of practitioners with the PVSG diagnostic criteria could be related to the lack of easy access and quick turnaround of the diagnostic tests listed in Table 1. Except for the determination of splenomegaly by the means of physical exam and widely available complete-blood count (CBC) that provides the determination of white-blood cell count (Table 1: diagnostic criterion B1), platelets (diagnostic criterion B2) along with hematocrit (HCT), all other diagnostic criteria are not widely available to the majority of clinicians. In addition, the determination of RBC volume involves radioactive chemicals, and the measurement of arterial gas saturation is an invasive and unpleasant test to perform. Consequently, further facilitation and improvement in our capacity to diagnose PV is needed. In particular, one would like to know what are the diagnostic yield of both CBC and the determination of splenomegaly by palpation, which are two widely available diagnostic tools in the diagnosis of PV.

In this paper, we present new rules, based on a reduced set of parameters, which can be used as a complement to the PVSG criteria. We have used standard classification techniques, artificial neural network (ANN) and support vector machine (SVM), for optimal selection of the parameters. Furthermore, we show that the projection of our dataset from the space defined by this reduced set of