UNIVERSITÀ DEGLI STUDI DI PADOVA
DIPARTIMENTO DI INGEGNERIA INDUSTRIALE
CORSO DI LAUREA MAGISTRALE IN INGEGNERIA CHIMICA E DEI PROCESSI INDUSTRIALI

Corso di Laurea Magistrale in
Ingegneria Chimica e dei Processi Industriali

DEVELOPMENT OF A MODEL-BASED APPROACH TO THE CHARACTERIZATION AND DIAGNOSIS OF TYPE 1 VON WILLEBRAND DISEASE

Relatore: Prof. Fabrizio Bezzo
Correlatore: Ing. Christopher Castaldello

Laureando: MATTEO BICEGO

ANNO ACCADEMICO 2017– 2018

Abstract

Von Willebrand disease (VWD) is the most common inherited bleeding disorder. It is caused by a deficiency of the von Willebrand factor (VWF), a protein involved in the haemostasis process. VWD is a very heterogeneous disease with many recognised typologies; type 1 VWD is the most common variant, covering almost 65% of the total cases. Type 1 VWD comprises several subtypes, making the disease diagnosis a complex task, which has to be carried out in specialized centres by expert medical practitioners. In many cases, DNA sequencing is required to obtain a certain diagnosis. However, about 35% of subjects affected by type 1 VWD do not carry any gene mutation, making even this test ineffective.

A pharmacokinetic model has been exploited to characterize each type 1 VWD subject and estimate useful pharmacokinetic indexes. These indexes and experimental measurements have been used to train support vector machine classifiers, proposing a model-based algorithm to support the diagnosis of type 1 disease variant. The results show that the implemented algorithm correctly classifies the subjects in about 90% of cases, thus representing a promising tool for reducing medical effort and improving patient lives.

Riassunto

La malattia di von Willebrand (VWD) è una delle più comuni malattie ereditarie che interessa il processo di coagulazione del sangue. La malattia è causata da un'insufficienza del fattore di von Willebrand (VWF), una proteina multimerica che permette l'adesione delle piastrine al vaso sanguineo danneggiato fermando il sanguinamento. La malattia è molto eterogenea ed in particolare lo è la variante più comune chiamata "tipo 1", caratterizzata da un'insufficienza quantitativa del VWF e rappresentativa del 65% circa dei casi di VWD.

La diagnosi è resa difficile dalla presenza di molti fattori, fisiologici ed ambientali, che influenzano il livello di VWF nel sangue. Inoltre, i sintomi, che tipicamente includono fenomeni emorragici, non sono specifici della malattia. Ciò comporta un lungo procedimento che comprende diversi test clinici per ottenere una diagnosi, che spesso richiede il sequenziamento del DNA per essere confermata.

Recenti studi hanno portato allo sviluppo di diversi modelli farmacocinetici per descrivere i fenomeni che interessano il VWF dopo il suo rilascio nel plasma. Tali modelli sono utilizzati sia per avere una migliore comprensione della malattia, sia per sviluppare dei metodi che possano essere di supporto al lavoro del personale medico specializzato durante la diagnosi.

In questo lavoro di ricerca, la categoria "tipo 1 VWD" è stata analizzata e caratterizzata attraverso l'utilizzo di un modello farmacocinetico in grado di descrivere i tre principali fenomeni che interessano il VWF: il rilascio, la proteolisi e l'eliminazione. I dati sperimentali sono stati forniti grazie alla collaborazione con l'azienda ospedaliera di Padova.

Successivamente, è stato possibile proporre un algoritmo che, utilizzando sia dati sperimentali sia informazioni ottenute dal modello farmacocinetico, attraverso tecniche di analisi dati e di apprendimento supervisionato per la classificazione, sia in grado di riconoscere se un soggetto sia affetto da VWD di tipo 1 o sia sano e, in caso sia malato, se la malattia sia riconducibile alla presenza di una mutazione nel gene che codifica il VWF oppure no.

I risultati mostrano una buona capacità del modello farmacocinetico di descrivere i soggetti malati di VWD di tipo 1, evidenziando le principali differenze tra le varie tipologie di soggetti malati caratterizzate da diverse tipologie di mutazioni.

L'algoritmo proposto, in fase di validazione, è in grado di classificare correttamente se un soggetto è malato, oppure sano, in oltre il 90% dei casi analizzati. Se il soggetto è malato, il modello di classificazione per determinare la presenza di una mutazione classifica erroneamente solo il 2.5% dei casi.

Table of contents

Introduction

The von Willebrand disease (VWD) is one of the most common inherited bleeding disorder and is caused by a deficiency of the von Willebrand factor (VWF). VWF is a multimeric protein, which helps the platelet adhesion to the damaged blood vessel during the haemostasis process. Qualitative deficiency of VWF characterizes variant type 1, which comprises many subtypes depending on presence and type of the VWF gene mutation.

Because of the difficulties to recognise the disease due also to the high heterogeneity of type 1 VWD, the disease diagnosis is a complex task, and several clinical tests have to be carried out by expert practitioners in specialized centres.

Several pharmacokinetic (PK) models have been developed to describe the phenomena involving the VWF after its release in blood plasma. PK models permit to characterize the subjects affected by the disease, thus providing a tool for a better disease comprehension; they can represent a starting point for proposing model-based approaches assisting medical doctors during the diagnosis of the disease.

In this Thesis a pharmacokinetic model able to quantify the VWF release, proteolysis and elimination phenomena, is used to characterized the subjects affected by type 1 VWD. Through experimental measurements and pharmacokinetic indexes derived from the model, an algorithmic approach supporting the diagnosis of type 1 VWD subtypes is proposed. The Thesis is structured as follows.

In Chapter 1, a general overview of the VWD is given, focusing on the description of the VWF synthesis and on the successive proteolysis mechanisms during the haemostasis process. The recognised typologies that characterized the VWD depending of the VWF deficiency are described. The medical protocol followed to achieve the diagnosis, together with the associated clinical test used in this Thesis, is then explained.

In Chapter 2, the pharmacokinetic model used to characterize the affected subjects is described. Particular attention is given to the parameters estimation activity. The mathematical methodologies adopted to formulate the diagnostic algorithm are eventually outlined.

In Chapter 3, the results obtained from the characterization of type 1 VWD subjects through the pharmacokinetic model are reported. In the first section, the results will be discussed focusing on the differences and similarity between the type 1 category and the healthy and the other VWD categories. The second section aims at describing the type 1 VWD subtypes in terms of relevant pharmacokinetic indexes.

In Chapter 4, the proposed model-based diagnosis algorithm, is explained. Classification results are reported and discussed.

Some final remarks and some advice for future work concludes the Thesis.

Chapter 1

The von Willebrand disease

In this chapter a general overview on the disease will be given. The role of von Willebrand factor (VWF) during the haemostasis process, the VWD categories and the clinical tests used for the diagnosis will be discussed.

1.1 Introduction to the von Willebrand disease

Von Willebrand disease (VWD) is one of the most common inherited bleeding disorder observed in humans. It was discovered in the 1926 by the Finnish physician Eric von Willebrand who took care of a 5 years old girl who showed severe bleeding disorder. What differentiate this bleeding disorder from haemophilia A was that it seemed not associated to muscles and joints bleeding. Both the girl's father and mother showed bleeding history, and several her sisters died after uncontrolled bleeding (Berntrop, 2007). A clear distinction from the haemophilia A was discovered only in the 1980's; persons who had VWD have a qualitative and/or quantitative von Willebrand factor (VWF) deficiency and normal FVIII gene while, person who had haemophilia A have an abnormal FVIII gene.
Based on the number of symptomatic patients seen at haemostasis centres, the VWD prevalence estimated values range between 0.0023 and 0.01 percent of the total population. Based, instead, on the symptoms estimations, the VWD prevalence values is between 0.6 and 1.3 percent (N.H.L.B.I., 2007). These discrepancies shows how difficult a certain VWD diagnosis can be.

1.2 The von Willebrand factor

The von Willebrand factor is a multimeric glycoprotein assembled from identical subunits each consisting of 2050 aminoacid residues and up to 22 carbohydrate side chains (Zaverio, 2007). VWF is found in circulating blood plasma, subendothelial cells and platelets and can reach a molecular weight as 20 million Dalton and a length up to 2 micrometers.
Von Willebrand factor plays a dual role during the haemostasis process. It helps the platelets adhesion to the injured vessel that otherwise, because of the high blood shear rate, would not be possible. Furthermore, circulating plasma VWF binds to FVIII preventing its cleavage. FVIII is a glycoprotein fundamental for the haemostasis process.

Levels of VWF are affected by several factors that includes age, ethnicity, blood group and hormones. VWF increasing during pregnancy, with aging and acute stress or inflammation. The VWF production rate probably is not affected by blood group, but VWF survival is reduced in persons who have type 0 blood.

1.2.1 The von Willebrand factor synthesis and secretion

The VWF gene is located at the tip of the short arm of chromosome 12 and it contains 52 exons. VWF is synthetized in two cell types; in the vascular endothelium by the endothelial cells and in bone narrow by the megakaryocytes. Endothelial synthesis involves the VWF cDNA translation that produce a precursor polypeptide referred to as pre-pro-VWF. The pre-pro-VWF and the propeptide together represent the pro-VWF, which is entirely composed of four type of repeating domains designated A through D. The order in which the different domains are arranged is reported in Figure 1.1. In the endoplasmic reticulum, VWF dimers are produced from the pre-pro-VWF monomers through disulphide bonds; the last 151 residues, following the C2 domain, are the only structure needed to dimerization. In the Golgi apparatus, concurrent with or soon after, the propeptide permits the VWF dimers associations, through the D domains, into high order polymers (Figure 1.2). (Zaverio 1999; Sadler, 2005).

The endothelial synthetized VWF can follow two different secretion pathway. A constitutive pathway, where the VWF is released as soon as its synthesis is completed, and a regulated pathway, where mature VWF is stored in the Weibel-Palade bodies, a rod-shaped organelles unique to endothelial cells. The VWF stored in Weibel-Palade bodies has a higher molecular mass than the constitutive VWF and it is released after stimulation by secretagogues.

The megakaryocytes VWF synthesis, because of the difficulties of platelets culturing, has not been studied in detail. Although definitive proof is still lacking, it is currently accepted that VWF is biosynthesized in megakaryocytes and endothelial cells in much the same way (Casonato, 2016). Furthermore, there are no doubt that all the VWF synthetized by megakaryocytes is stored in α-ganules of platelets and is released only after platelets activations during the haemostasis process (Zaverio, 1999).

The VWF stored in α-ganules, similarly to the one stored in Weibel-Palade, is composed of the largest multimers, showing an enhanced haemostatically action.

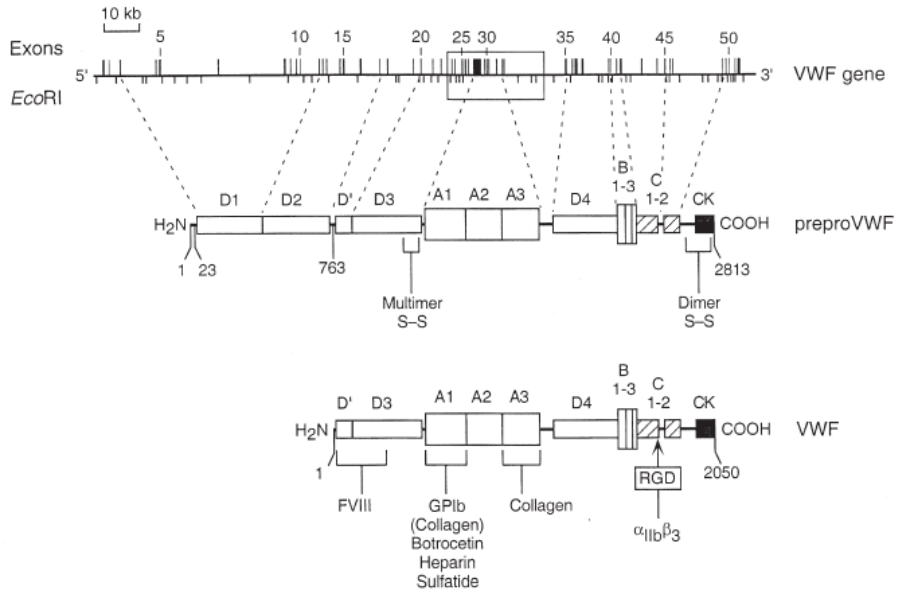


Figure 1.1. Gene and VWF structures. The exons that codify each pre-pro VWF domains are highlighted (Sadler, 2000).

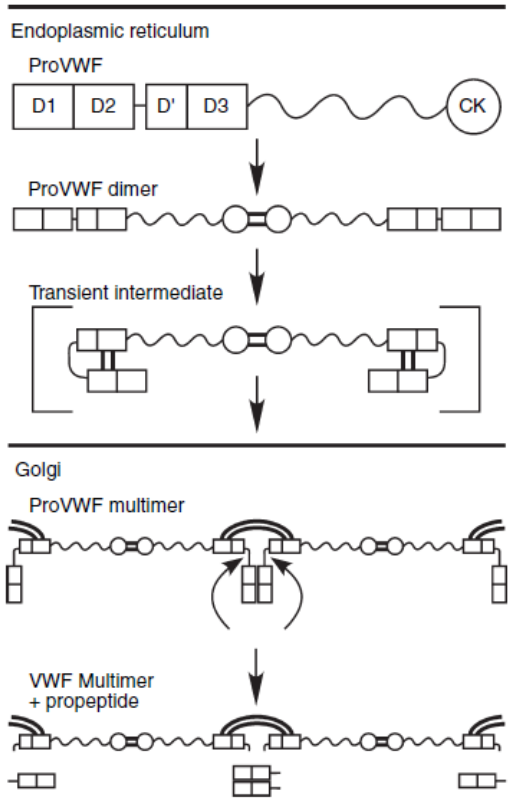


Figure 1.2 VWF dimerization process in the endoplasmic reticulum and VWF synthesis in the Golgi apparatus (Sadler, 2005).

The VWF circulating in plasma is essentially originated by the endothelial cells and is secreted following the constitutive pathway. The proteolytic action of ADAMST13 metalloproteinase enzyme at the bond between Tyr842 and Met843 in the A2 VWF domain convert the large multimers into smaller ones. The A2 domain is exposed to the ADAMST13 actions when VWF multimers are subjected to sufficient blood shear rate (Sadler, 2005). Proteolytic actions is necessary to prevent the thrombosis formation. Platelet VWF is more resistant to the ADAMSTS13 proteolytic action, in fact the VWF subunits fragments produced by plasma VWF degradation are not found in platelets VWF (Sadler, 2005).
VWF multimers are then cleared with a half-life of 12-20 h by a mechanism that not strongly depend on multimers size.

1.2.2 The haemostasis process and the role of VWF

Haemostasis is the sum of the physiological processes that allows bleeding stops at the injury site, maintaining normal blood flow in the circulation. Four different steps characterize haemostasis:

- Vascular spasm;
- Primary haemostasis;
- Secondary haemostasis;
- Fibrinolysis.

The vascular spasm is necessary to reduce the blood flow in the injured site and is caused by the endothelins releasing from the endothelium. Endothelins are amino-acid peptides with a powerful vasoconstriction action.
Primary haemostasis involves the platelets adhesion to the damaged vessel wall and the subsequent growth of the haemostasis plug due to the platelet-platelet interactions. VWF, which is released from the sub-endothelium, plays a central role during the primary haemostasis process; VWF A3 domain contains the main binding site for collagen. In a normal shear rate blood flow (1000 s^{-1}), only binds between GP Iba glycoprotein present on the platelets surface and immobilized VWF A1 domain can initiate the platelets adhesion to the vessel wall (Figure 1.3). The platelet adhesion in low shear rate flow (500 s^{-1}) is not affected by the absence of VWF.
The VWF and platelets bond affect the platelets structure, which released adenosine diphosphate and thrombin; this form of platelets is called activated form (Figure 1.4). The change in form of platelets expose the integrin aIIbβ3, which become the binding site for adhesive ligand (mainly fibrinogen and VWF) for the additional attachment of platelets (Zaverio, 2007).

Secondary haemostasis consists of a cascade process that culminates with the activation of fibrin (Figure 1.5). The cascade process can be initiate by two events, which characterized two pathway, the intrinsic pathway and the extrinsic pathway.

The extrinsic pathway is activated only after blood vessel injury after the release of tissue factor (TF) and produce thrombin, which cleaves fibrinogen to generate insoluble fibrin. Furthermore, thrombin activates factor XI that initiates the intrinsic pathway. Intrinsic pathway can be activated by blood contacts with extracellular collagen. Factor VIII participates in intrinsic pathway and is present in plasma, bounded to VWF to prevent its degradations. The VWF D'-D3 domains form the site that binds FVIII.

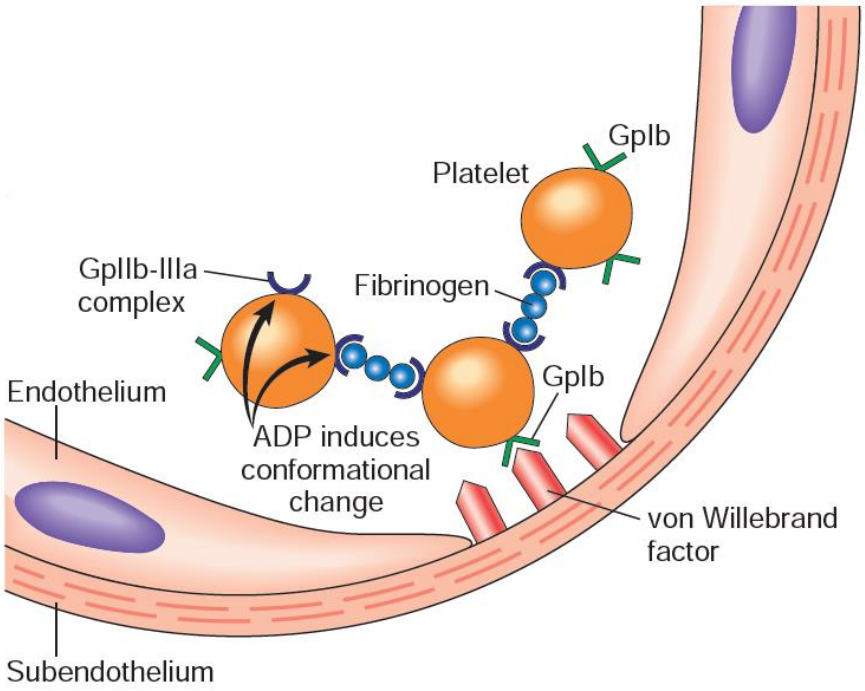


Figure 1.3. Platelets adhesion and aggregation. The binding sites between platelets and VWF are highlighted (Kumar, 2015).

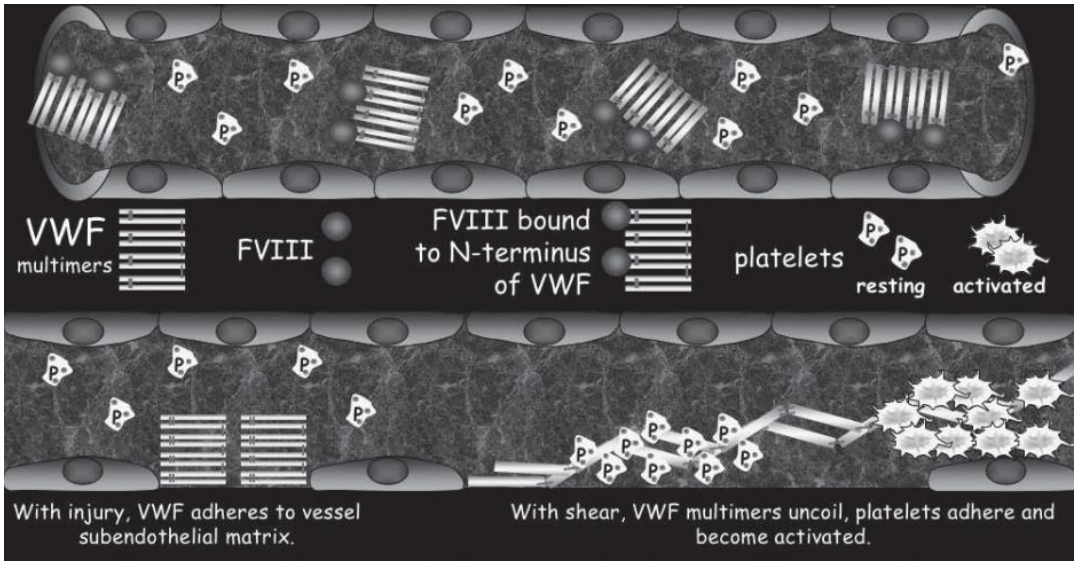


Figure 1.3. Primary haemostasis process: VWF vessel adhesion, uncoiling and platelets activations (N.H.B.L.I., 2007).

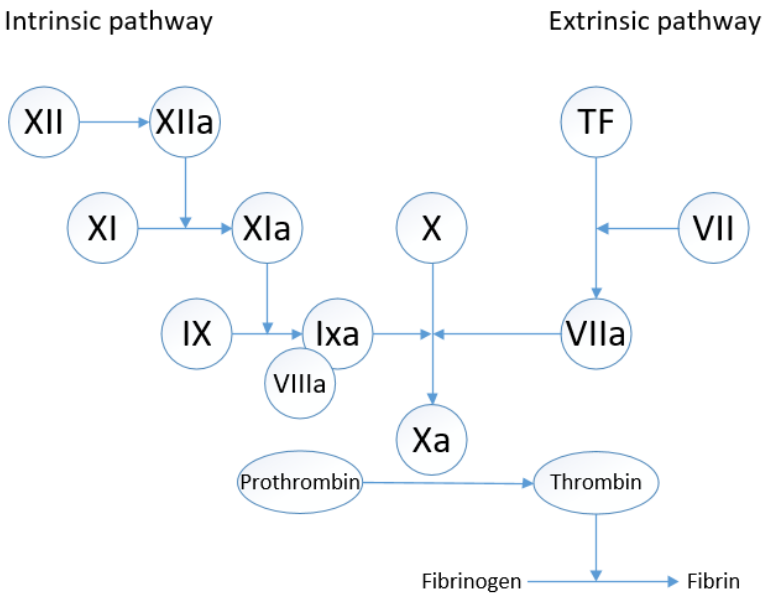


Figure 1.4. Secondary haemostasis process scheme (Adapted from N.H.B.L.I., 2007).

Finally, fibrinolysis aims to dissolve the blood clots during the process of wound healing and stops the cloth formation in the healthy blood vessels.

1.3 Classification of von Willebrand disease

Von Willebrand disease appears in several forms and its classification is intended to be clinical relevant to its treatment. The 1994 classification reserved the VWD designation for disorders caused by VWF gene mutations, but in 2006 this criterion has been dropped because in practice it is verifiable for only a small fraction of patients (N.H.L.B.I., 2007).
VWD is classified into three major categories: type 1 is characterized by partial VWF deficiency, type 2 is characterized by qualitative deficiency and type 3 is characterized by total deficiency.

1.3.1 Type 1 VWD

Type 1 VWD is the most common form of VWD, accounting for about 75 % of the total VWD affected subjects. The level of VWF in plasma is low but it mediates platelets adhesion and binds FVIII normally.
VWF is a multimer and a single mutant subunits might impair the intracellular transport of VWF dimers or the multimers secretion, decreasing the VWF levels without affects the multimers distribution. Sometimes, reduced secretion is not enough to justify low VWF concentration, suggesting that an enhanced clearance might contribute (Sadler, 2005).
VWD type 1 diagnosis is a difficult task. Some cases shows exceptionally low VWF levels, repeated and serious bleeding disorders and have dominant negative mutations that interfere with the intracellular transport of dimeric pro-VWF or promote its clearance from the circulation. The subjects affected by severe type 1 usually have VWF levels lower than 20 IU/dL while the VWF normal range is between 50 and 200 IU/dL.
Levels of VWF just below the normal range (30-50 IU/dL) pose a problem for diagnosis. The broad VWF normal range significantly overlaps the VWF levels of mild VWD type 1. Moreover, VWF deficiency can cause bleeding, but bleeding has many causes so it identifies a risk factor for bleeding but does not mandate a diagnosis of type 1 VWD (Sadler, 2003). Furthermore, about the 35 % of the mild type 1 VWD affected subjects have no VWF mutations (Lillicrap, 2007).
For all these reasons, a certain diagnosis is often difficult to obtain, and most diagnosis of type 1 VWD are false positives. The consequences of VWD misdiagnosis are not necessary benign. Patients may be exposed to risky, expensive and useful treatments, while the real cause of symptoms are untreated (Sadler, 2003).

1.3.2 Type 2 VWD

Type 2 VWD is characterized by a qualitative VWF deficiency. There are many variants of type 2 VWD due to the different reasons that cause the anomaly, the following variants are the most known.

Type 2A VWD. This variant shows a qualitative deficiency refers to a decreased VWF-dependent platelet adhesion because the proportion of large VWF multimers is decreased. Levels of VWF:Ag and FVIII may be normal or modestly decreased. Type 2A VWD may be caused by mutations that interfere with the assembly or secretion of large multimers or that increases the VWF multimers proteolytic degradations. Mutations are located in the VWF A2 domain.

Type 2B VWD. For this variant an increase platelet-VWF affinity occurs. The circulating platelets are coated with mutant VWF, which may prevent the platelets adhesion to the injured site. Therefore, large, functional VWF multimers are cleaves by the ADAMST13 actions. Mutations are located within or adjacent to the VWF A1 domain and involves a VWF conformational change, which enhances the VWF-platelet binding.

Type 2N VWD. This variant shows an impaired FVIII binding. FVIII is so exposed to degradation. Mutations are located between domains D' and D3. Because the low FVIII levels, type 2N VWD can be masquerades from haemophilia A and specific assay are required to prevent misdiagnosis. The capital n stand for Normandy, where the firsts cases where discovered.

Type 2M VWD includes variants with decreased VWF-dependetn platelet adhesion not caused by the absence of high molecular weight VWF multimers. TYPE 2m mutations does not affect the VWF impairing but reduce the VWF interaction with platelet or with connective tissue (collagen). Mutations are located in A1 domain, where they interfere with platelet GBIb.

Type Vicenza VWD is a variant with high discrepancy between the plasma and platelets VWF. The plasma VWF level is low but the VWF multimers are ultra-large, meaning that the metalloproteinase ADAMS-13 is not able to proteolysis VWF. The low VWF level in plasma can be explained by an enhanced clearance while normal level of platelets VWF suggests a normal synthesis. For these reasons, depending of interpretations of laboratory tests, sometimes types Vicenza are classified under type 1 or under type 2M. Mutations for type Vicenza are specific of Arg1205His (N.H.L.B.I., 2007). This VWD subtypes is called Vicenza because of the place where the specific mutation was found.

1.3.3 Type 3 VWD

Type 3 VWD is characterized by undetectable VWF protein and activity. FVIII level are usually very low (< 9 IU/dL). Mutations are distributed throughout the VWF gene and most are unique to the family in which they were first identify. Nonsense and frameshift mutations are the most common causes to type 3 VWD, but also large deletions and missense mutations can do so.

1.3.4 Acquired VWD

Acquired VWD syndrome (AVWS) refers to defects in VWF concentration, structure or function not inherited but that are consequences of other medical disorder. The VWF multimers distribution may be normal, but often shows a reduction in the large multimers. The main AVWS causes are autoimmune clearance or inhibition of VWF, increased shear-induced VWF proteolysis or increased bindings between VWF and platelets or collagen.

1.4 Diagnosis and VWD detection

The VWD diagnosis and the correct types detection are obtained taking into consideration the personal bleeding history of the patient with some information about the presence of a family history of bleeding events and the laboratory tests to detect quantitative o qualitative VWF deficiency. Because of the numerous VWD types and the high heterogeneity that characterized VWD, an algorithm reported in Figure 1.6 was proposed to prevent the misdiagnosis.
The initial clinical assessment is focus on his personal and family history of excessive bleeding disorders. There must be identify the bleeding sites, severity, duration and type of injury associated to bleeding. This first evaluation is challenging, in fact mild bleeding disorder are very common in healthy populations. In addition, VWD cause bleeding, but bleeding is not specific related to VWD. A positive family history is useful to identify persons who are likely to have VWD, but in cases of mild VWD, such family history could not be present.

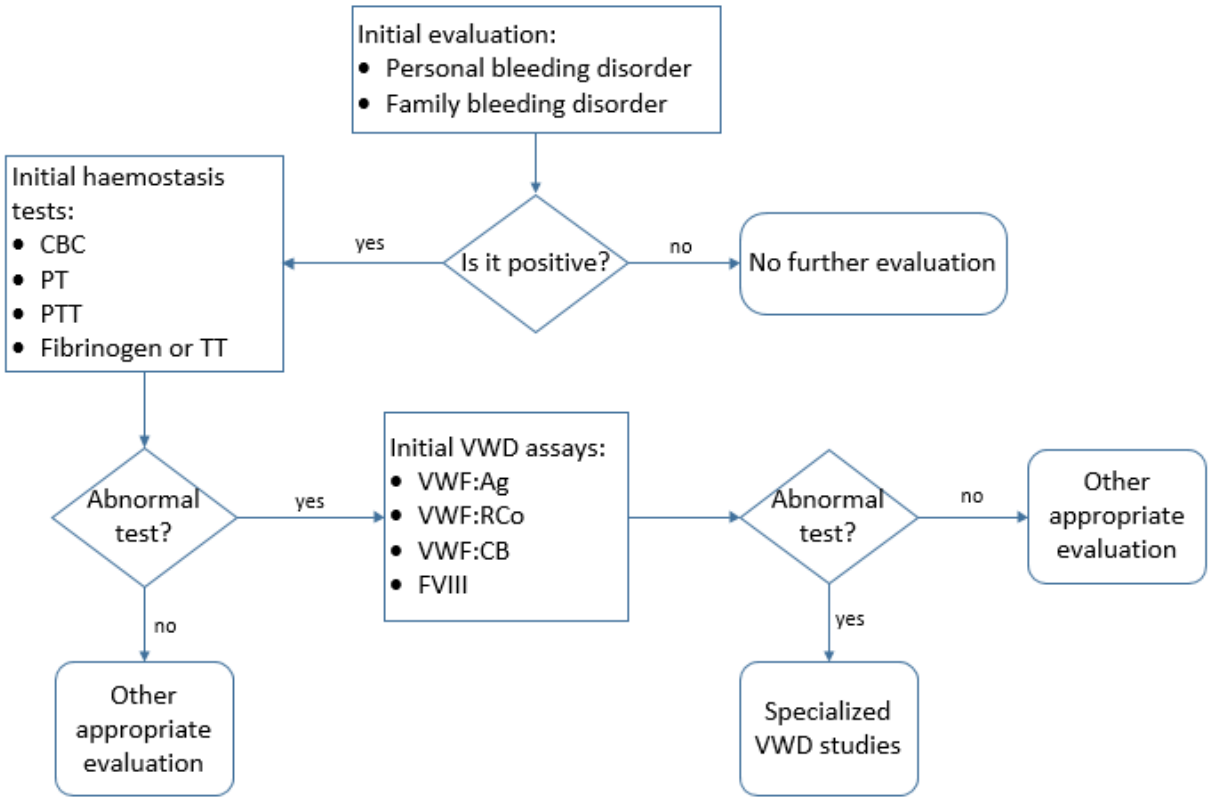


Figure 1.1. Diagnosis algorithm used for VWD diagnosis (Adapted from N.H.B.L.I., 2007).

The common initial haemostasis laboratory tests are:

- Complete blood count (CBC); a test that determine the concentrations of the main cells presents in the blood (platelets, white blood cells, red blood cells);
- Prothrombin time (PT); a test that measures the time needed to cloth formation by the extrinsic pathway of haemostasis that involves factor I, II, V, VII and X.
- Partial tromboplastin Time (PTT); a tests that measures the time needed to cloth formation by the intrinsic pathway;
- Fibrinogen; a test used to determine the fibrinogen concentration. Fibrinogen is enzymatically converted to fibrin by thrombin;
- Thrombin Time (TT); a test that measures the clotting time and indicates the conversion of fibrinogen to fibrin.

If one or more of these tests results abnormal, a more specific set of tests is needed to diagnosis the VWD. The initial tests for VWD includes measures of the VWF amount present in plasma, the function of VWF protein and the VWF ability to maintain FVIII in blood circulation.
In order to obtain a diagnosis and detection of VWD sub-types an important test is the DDAVP. DDAVP test involves the vasopressin (1-deamino-8D-arginine vasopressin) administration and the collection of blood samples following a precise time schedule (0, 15, 30, 60, 120, 180, 240,

360, 480, 1440 min). Vasopressin induce the VWF release stored in the Weibel-Palade bodies, which is then exposed to the proteolytic actions of ultralarge multimers and then undergoes to its clearance. Each blood sample is then subjected to VWD specialized laboratory tests.

1.4.1 Clinical tests

VWF:Ag measures the concentration of VWF protein presents in plasma and it is expressed in International Units (IU) per volume. The method used is based on enzyme-linked immunosorbent assay (ELISA). The reference curve is construct using a pool of normal washed platelets (N.H.L.B.I., 2007).

VWF:RCo measures the VWF ability to interacts with normal platelets. Ristocetin antibiotic causes the binds between VWF and platelets resulting in platelets clumps that is then removed from circulation. Ristocetin is a valid test for VWD but it has been removed because it caused thrombocytopenia (N.H.L.B.I., 2007).

VWF:CB measures the VWF ability to binds with collagen and it is expressed in International Units (IU) per volume. The collagen binding assay is dependent on VWF multimeric size, with the larger multimers that binds more than the smaller. The VWF:CB is used for both VWD detection and VWD discrimination among VWD types (N.H.L.B.I., 2007).

VWF:FVIII binding measures the VWF ability to binds factor VIII. VWF bind FVIII so to prevent its cleavage. It is used to distinguish between VWD and haemophilia A, and to diagnose type 2N VWD (N.H.L.B.I., 2007).

Platelet VWF measures the VWF concentration in platelets. Platelet VWF is a measure of the VWF synthesis quality; it is accepted that platelet VWF is synthetized in much the same way than the plasma VWF, but it is more resistant to the ADAMST13 proteolytic action and so it shows an enhanced haemostatic action (Casonato, 2016).

DNA sequencing is a DNA test. For VWD type 2, the cDNA mutations are directly related to each subtypes (2M, 2B, 2N, 2A), while for VWD type 1 this test could be useless. In most person who have type 1 VWD, the genetic mutations have not been established (N.H.B.L.I., 2007). DNA sequencing is not widely available.

1.4.2 Model-based VWD Characterization and diagnosis

Because of the difficulties related to the VWD diagnosis and characterization discussed in the previous chapter several pharmacokinetic models have been proposed. The model proposed by Galvanin (2014) aims of describing the phenomena that occurs after the DDAVP administration; from the VWF multimers release from the subendothelium, passing through

their proteolysis to their clearance from the circulating blood plasma. A model able to describe the patient's biological pathway in a reliable way is a powerful tool in the VWD diagnosis and characterization; furthermore it may permit the identification of a more effective therapies providing a better life for the patient and saving medical efforts.
Recent works (Ferrari *et al*, 2018) proposed a simplified Galvanin pharmacokinetic model which is able to predict the VWF:Ag and VWF:CB levels. The model identifiability involves the DDAVP test (as will be discuss in the next Chapter) with a precise time schedule sampling activity. The VWD pharmacokinetic model is tested through its identifiability from experimental measures and assessing its ability to represent the predicted variables. A pharmacokinetic model that well represent the phenomena involved permits to obtain some information about the system described that are not directly shown in experimental measures and allows a deeper process knowledge, which can be used both to facilitate the medical work and to improve the patient life. Appling an advance model-based technique it is possible to reduce the DDAVP time required to model identification, while through some pharmacokinetic indexes is possible to well characterized each subject allowing the development of a model-based technique to assist the medical decision making task in the diagnosis fields (Castaldello et al, 2017).

1.5 Thesis objective

The VWD type 1 is the most common VWD, covering about the 65% of the total VWD cases and because of its high heterogeneity is the most difficult to diagnose. In particular, the diagnose of mild type 1 VWD is very challenging and also the genetic test is often useless. The simplified pharmacokinetic model proposed by Ferrari *et al* (2018), showed good results in terms of identifiability and predicted responses and it was tested on the VWD type 2A, 2B and Vicenza. The first goal of this thesis work is to test the simplified PK model on the VWD type 1 obtaining a pharmacokinetic characterization of the entire pull of patients. The characterization can be made through the model parameters or using some PK indexes; both will be described in the next Chapter.
The information given by the PK model will then be used to propose a procedure for VWD diagnosis.

Chapter 2

Mathematical methods

Pharmacokinetic models have been developed to describe the VWF processes that occur in the organism following the DDAVP administration. In this chapter, the PK model used is described and its mathematical formulation will be provided. A mathematical background for the parameters estimation activity will be given and some tests used in this Thesis to assess the estimation quality will be discussed. The development of a model-based classification protocol involves some data-analysis and pattern recognition techniques; in this chapter the main techniques used will be discussed and their mathematical formulation provided.

2.1 Simplified Pharmacokinetic model of von Willebrand disease

After the vasopressin (DDAVP) administration, three principal phenomena occur involving the VWF multimers:

1. Release of VWF multimers;
2. VWF proteolysis to smaller multimers by ADAMST13 action;
3. Clearance of VWF (independently of multimers size).

Several pharmacokinetic models has been proposed (Galvanin *et al*, 2014) to describe the courses of plasma VWF:Ag and VWF:CB levels after the vasopressin administration.
A reliable pharmacokinetic model of VWD should have these features (Galvanin *et al*, 2014):

- It should represent the real physiological pathways involved in the time evolution of VWF concentration;
- its parameters should be easily identifiable from clinical tests available;
- it should be sufficient flexible to well-describe the single patient;
- it should represent the multimers distribution in time.

The model adopted in this Thesis work is the one proposed by Ferrari *et al*, (2018) which is a simplification of the model originally developed by Galvanin *et al* (2014).
The model, sketched in Figure 2.1, comprises two compartment and it is designed based on the following assumptions (Ferrari *et al*, 2018):

- at the basal state only high (H) and low (L) molecular weight multimers are presents;

- the release of ultra-large (UL) and high (H) molecular weight multimers is consequence of vasopressin administration;
- the elimination is considered independent of the multimers size;
- VWF:Ag measures the total amount of VWF, comprising UL, H, and L multimers;
- VWF:CB measures the amount of UL and H multimers size.

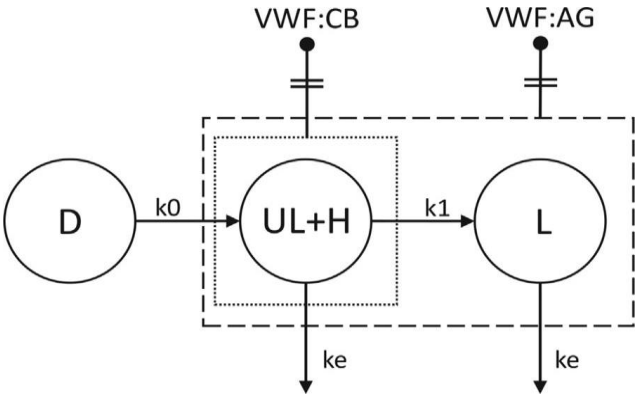


Figure 2.1. Pharmacokinetic model structure with the two compartments and the location of VWF:Ag and VWF:CB measurements (Adapted from Ferrari *et al*, 2018).

The model is described by the following set of differential and algebraic equations:

$$\frac{dx^{UL+H}}{dt} = k_0 D e^{-k_0(t-t_{max})} - k_1(x^{UL+H} - x_b^{UL+H}) - k_e(x^{UL+H} - x_b^{UL+H}) , \quad (2.1)$$

$$\frac{dx^{L}}{dt} = k_1 (x^{UL+H} - x_b^{UL+H}) - k_e (x^{L} - x_b^{L}) , \quad (2.2)$$

where x^{UL+H} and x^{L} are the number of UL+H and L multimers units respectively and the subscript *b* stands for the basal condition. The amount of VWF released is represented by the *D* parameter while k_0 is the parameter that quantifies the release rate. k_1 and k_e are respectively the proteolysis and elimination rate. t_{max}, called lag time, is the time of the maximum response. The VWF:Ag and VWF:CB concentration measurements are respectively related to the multimers units as:

$$y^{Ag} = \frac{x^{UL+H} + x^{L}}{V_d} , \quad (2.3)$$

$$y^{CB} = \frac{x^{UL+H}}{V_d} , \quad (2.4)$$

where the distribution volume V_d is approximated using the body weight *BW* following the equation (Menache et al., 1996):

$$V_d = 0.4\, BW . \quad (2.5)$$

In order to take into account the different affinity of multimers to collagen between different VWD type, a correction factor *k* is introduced:

$$y^{CB'} = k\, y^{CB}\, \frac{y_b^{Ag}}{y_b^{CB}}. \tag{2.6}$$

The conversion from multimers units to concentration is expressed in the following equations:

$$x_b^{UL+H} = y_b^{CB}\, V_d\ , \tag{2.7}$$

$$x_b^{L} = y_b^{Ag}\, V_d - x_b^{UL+H}\ . \tag{2.8}$$

The parameters set **θ** [theta] that has to be estimated comprise seven parameters obtained after a riparametrization:

$$\boldsymbol{\theta} = [k_0, k_1, k_e, D/t_{max}\,, k, y_b^{CB}, t_{max}]. \tag{2.9}$$

The parameters estimated set permit to define some pharmacokinetic indexes which are commonly used to the subjects characterization as the *Q* amount of VWF multimers released, the *CL* blood volume treated for unit of time and weight and the *R* ratio between the VWF:CB and VWF:Ag at the basal state. These PK are defined by the following equations:

$$Q_{re}^{UL+H} = Q_{re} = \frac{1}{BW}\int_0^{\tau} k_0\, D\, e^{-k_0\,(t-t_{max})}\, dt\ ; \tag{2.10}$$

$$CL^{UL+H} = V_d\, k_e\, \frac{1}{BW}\ ; \tag{2.11}$$

$$R = \frac{vWF{:}CB_b}{vWF{:}Ag_b}\ ; \tag{2.12}$$

where the subscript *re* stand for released and τ is the experiment time.

2.1.1 Parameters estimation

A model *M* can be represented by a system of differential and algebraic equations (DAEs) as:

$$M{:}\begin{cases} f\big(\dot{\boldsymbol{x}}(t), \boldsymbol{x}(t), \boldsymbol{u}(t), \boldsymbol{w}, \widehat{\boldsymbol{\theta}}, t\big) = 0 \\ \widehat{\boldsymbol{y}}(t) = \boldsymbol{h}(\boldsymbol{x}(t)) \end{cases}, \tag{2.13}$$

where:

- $\boldsymbol{x}(t) \in \Re^{N_x}$ is the vector of time-dependent state variables;
- $\dot{\boldsymbol{x}}(t) \in \Re^{N_x}$ is the derivative vector of time-dependent state variables;
- $\boldsymbol{u}(t) \in \Re^{N_u}$ is the set of the time-dependent control variables;
- $\boldsymbol{w}(t) \in \Re^{N_w}$ is the vector of time-invariant control variables;
- $\widehat{\boldsymbol{\theta}} \in \Re^{N_\theta}$ is the set of unknown parameters to be estimated;
- t is the time of the experiment;

- $\hat{\boldsymbol{y}}(t)\ \epsilon\ \Re^{N_y}$ is the vector of predicted variables.

When a model has been selected among a set of candidate, its identification is due to the parameters identification through some available experimental measures. The aims of the parameters estimation activity are to obtain the parameters values that maximise the model capability to predict the measured response and to achieve a satisfactory statistically parameters estimation.
An estimator Φ^{PE} can be defined as:

$$\widehat{\boldsymbol{\theta}} = \Phi^{PE}(\boldsymbol{y}): \Re^{N_y} \rightarrow \Re^{N_\theta} \tag{2.14}$$

where $\widehat{\boldsymbol{\theta}}$ is the set of estimated parameters.
The quality of an estimator is assessed by the quality of the estimated parameters in terms of accuracy and precision. Estimated parameters accuracy is related to the nearness to its true value; in general this feature cannot be verified because the true values of the model parameters are unknown. Parameters precision is achieved with the minimum dispersion around the estimated values.
A satisfactory estimation result should be a vector of parameters having the minimum variance and providing the minimum deviation between the predicted response $\hat{y}$ and the measured variables *y*. This can be obtain by minimizing the residuals quantity r_{ij} defined as:

$$r_{ij} = ||y_i(t_j) - \hat{y}_i(t_j)|| \qquad i = 1, \dots, N_y,\ j = 1, \dots, N_{sp}, \tag{2.15}$$

where *t* is the time of the experimental measure, N_y is the response number and N_{sp} is the experimental measures number.
The measurement values are affected by measurement uncertainty and for this reason, an associated probability distribution function (PDF) should be defined. The most widespread univariate PDF is called normal distribution function and is describe by the following equation:

$$p(x) = \frac{1}{\sqrt{2\,\pi\,\sigma^2}}\ e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}, \tag{2.16}$$

where μ is the mean and σ is the standard deviation.
If the model used to describe the system is correct and only measurement errors affects the system responses, it is reasonable to model the residues PDF as normally distributed with zero mean and a certain standard deviation σ_{ij}. Considering that the residues PDF is function of the set of parameter θ since $\hat{y}_i(\theta)$, the joint probability function, assuming the residuals as uncorrelated, is defined by the likelihood function:

$$L(\boldsymbol{\theta}) = \prod_{i=1}^{N_{exp}} \prod_{j=1}^{N_m} \frac{1}{\sqrt{2\,\pi\,{\sigma_{ij}}^2}}\, e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma_{ij}}\right)^2} . \tag{2.17}$$

The minimization of the residual from equation (2.15) can be reformulate through the likelihood maximization, thus the parameters estimation problem can be write as (Bard, 1974):

$$max_\theta\{L(\boldsymbol{\theta})\} = max_{\boldsymbol{\theta}} \left\{ \prod_{i=1}^{N_{exp}} \prod_{j=1}^{N_m} \frac{1}{\sqrt{2\,\pi\,{\sigma_{ij}}^2}}\, e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma_{ij}}\right)^2} \right\}. \tag{2.18}$$

The use of equation (2.18) approach rather than equation (2.15) has many advantages: it takes into account the measures uncertainly and also permits to obtain some a-posteriori statistics useful to assess the quality of the estimation.

To achieve the parameters estimation using the maximum likelihood method, the following elements are required (Galvanin, 2010):

- a model with its initial condition;
- an initial guess for the entire set of parameter;
- a set of design vectors that define the experimental setting;
- a **y** set of experimental data;
- information about the measurement system expressed by the variance-covariance matrix $\boldsymbol{\Sigma}_y$.

If no information about $\boldsymbol{\Sigma}_y$ is given, is possible to use a variance model in the form of the following equation:

$$\sigma_{yj} = \omega_j^2 (\hat{y}_j^2)^{\gamma_i} \tag{2.19}$$

where the γ and ω are the two model parameters which are defined depending on the variance model type applied. The possible values are summarized in Table 2.1.

Table 2.1 Variance model parameters values depending on the variance model choose.

Expected variance model	γ	ω
Constant variance	0	Fixed *a-priori* or estimated
Constant relative variance	1	Fixed *a-priori* or estimated
Heteroschedastic	Fixed *a-priori* or estimated	Fixed *a-priori* or estimated

2.1.2 Information content analysis

When a model has been selected among a set of candidate, its identification can be divided in two steps: the *a priori identifiablility* and the *a posteriori identifiability*. The *a priori identifiability* activity aims of verifying if the model parameters can be estimated in a noise-free and disturbance-free conditions.
The *a priori identifiability* condition is verified if given two different parameters set $\boldsymbol{\theta}$ and $\boldsymbol{\theta}^*$ the following condition:

$$M(\boldsymbol{\theta}) = M(\boldsymbol{\theta}^*) \Rightarrow \boldsymbol{\theta} = \boldsymbol{\theta}^* \tag{2.20}$$

is verified and so the model responses are the same if the parameters set are equal. If equation (2.20) is always satisfy the model is *Structurally Globally identifiable* (SGI), while it is *Structurally Locally identifiable* (SLI) if there exist a $\boldsymbol{\theta}$ neighbour that satisfy condition (2.20).
A priori identifiability is a necessary condition to guarantee successful parameters estimation using noise-affected real data which characterized the *a posteriori identifiability.*
Sensitivity analysis is the study of how a variation in one parameter affects the model outputs and it is used to assess the *a posteriori identifiability.*
Local and global sensitivity analysis have been proposed, for attaining two goals:

- to detect the most influential parameters affecting the system responses;
- to analyse the information behaviour for a given set of experimental conditions.

Sensitivity analysis is carry out by perturbing the PK parameters (1%) and observing how it affects the model responses. Mathematically this can be expressed as:

$$q_i^{Ag} = \frac{\hat{y}_{Ag}(\theta_i') - \hat{y}_{Ag}(\theta_i)}{\theta_i' - \theta_i} \qquad i = 1, \dots, N_\theta \tag{2.21}$$

$$q_i^{CB} = \frac{\hat{y}_{CB}(\theta_i') - \hat{y}_{CB}(\theta_i)}{\theta_i' - \theta_i} \qquad i = 1, \dots, N_\theta \tag{2.22}$$

where θ_i' and θ_i represent the perturbed and original set of parameters while N_θ is the number of model parameters.
Let N_y number of model responses, the $N_y \times N_\theta$ matrix of local sensitivities $\boldsymbol{Q}$ is represented by the following equation:

$$\boldsymbol{Q}(t) = \begin{bmatrix} q_{1,1}(t) & \cdots & q_{1,N_\theta}(t) \\ \vdots & \ddots & \vdots \\ q_{N_y,1}(t) & \cdots & q_{N_y,N_\theta}(t) \end{bmatrix} = \begin{bmatrix} \frac{\partial \hat{y}_1(t)}{\partial \theta_1} & \cdots & \frac{\partial \hat{y}_1(t)}{\partial \theta_{N_\theta}} \\ \vdots & \ddots & \vdots \\ \frac{\partial \hat{y}_{N_y}(t)}{\partial \theta_1} & \cdots & \frac{\partial \hat{y}_{N_y}(t)}{\partial \theta_{N_\theta}} \end{bmatrix} \tag{2.23}$$

which can be evaluated for each r-*th* measured response at each sampling time through the $n_{sp} \times N_\theta$ matrix $\boldsymbol{Q}_r$ represented in equation (2.24).

$$\boldsymbol{Q}_r = \begin{bmatrix} \left.\frac{\partial y_r}{\partial \theta_i}\right|_{t_1} & \cdots & \left.\frac{\partial y_r}{\partial \theta_{N_p}}\right|_{t_1} \\ \vdots & \ddots & \vdots \\ \left.\frac{\partial y_r}{\partial \theta_i}\right|_{t_{n_{sp}}} & \cdots & \left.\frac{\partial y_r}{\partial \theta_{N_p}}\right|_{t_{n_{sp}}} \end{bmatrix}. \tag{2.24}$$

The analysis of time profile of local sensitivities shows the dynamic behavior of the parametric system, providing useful information about the parameters estimation. Considering a single response model with four parameters, looking at the sensitivities profile shown in Figure 2.2, it is possible to note that parameters q_{11} and q_{14} shows great sensitivity on the model response, while for parameter q_{12} the sensitivity is null. About the parameters estimation activity it is reasonable to think that parameter q_{12} will be hardly identifiable since it does not affects the model response, while parameters q_{11} and q_{14} seem to be easily identifiable but they are strongly anti-correlated.

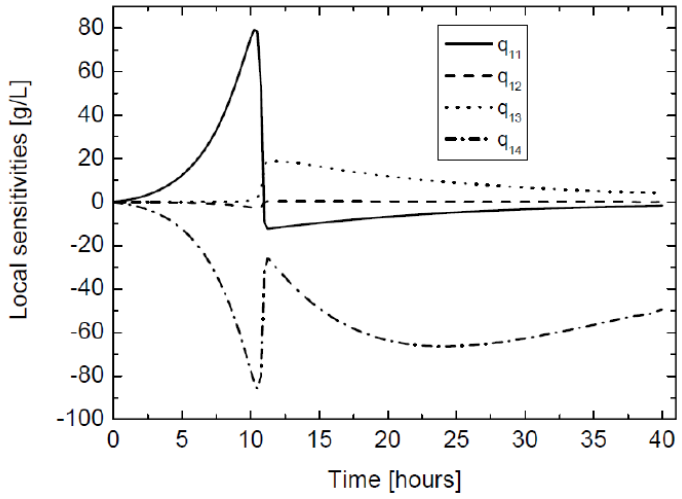


Figure 2.2 Dynamic sensitivities profile for a model with four parameters and a single response (Galvanin, 2010)

The dynamic sensitivity matrix is strictly correlated to the information foreseen by the model identification. For a multiple input, multiple output (MIMO) dynamic system the information carried by the model can be defined as:

$$\boldsymbol{I}(\boldsymbol{\theta}, t) = \left[\boldsymbol{Q}'^{T}(t)\, \boldsymbol{\Sigma}_y^{-1}\, \boldsymbol{Q}^{T}(t)\right] \tag{2.25}$$

Where $\boldsymbol{\Sigma}_y$ is the $N_y \times N_y$ variance-covariance measurement matrix and $\boldsymbol{I}$ is the information matrix. Under the Zullo hypothesis (Zullo,1991) an information matrix discrete form can be formulated as:

$$\boldsymbol{H}_\theta(\boldsymbol{\theta}, \boldsymbol{\varphi}) = \sum_{k=1}^{N_{sp}} \sum_{i=1}^{N_y} \sum_{j=1}^{N_y} \tilde{\sigma}_{ij|_k}\, \boldsymbol{Q}_{i|_k}^{T}\, \boldsymbol{Q}_{j|_k} + \boldsymbol{H}_\theta^0 . \tag{2.26}$$

The estimated model parameters variance-covariance matrix $\boldsymbol{V}_\theta$ is the inverse of the dynamic information matrix $\boldsymbol{H}_\theta$:

$$\boldsymbol{V}_\theta(\boldsymbol{\theta},\boldsymbol{\varphi}) = [\boldsymbol{H}_\theta(\boldsymbol{\theta},\boldsymbol{\varphi})]^{-1} = \left[\sum_{i=1}^{N_y}\sum_{j=1}^{N_y}\tilde{\sigma}_{ij}\,\boldsymbol{Q}_i^T\,\boldsymbol{Q}_j\right]^{-1}. \quad (2.27)$$

Model parameters variance-covariance matrix $\boldsymbol{V}_\theta$ is used to assess the parameters quality.

2.1.3 Estimation quality

The parameters estimation activity aims of solving model equations (2.13) minimizing the objective function (2.15); statistical tests are essential to verify the estimated parameters quality.
A satisfactory parameters estimation should have the following features (Emery, 2001):

- accuracy: the estimate parameters has to capture the information embedded by the experimental measures, rejecting the noise and disturbances effects;
- precision: the parameters uncertainty has to be minimized.

The parameters estimated precision is strictly related to the parameters variance-covariance matrix defined by equation (2.27).
If the parameters are assumed to be normally distributed, the t-test is a powerful test to understand if parameters are well estimated or not. Defining the confidence interval as:

$$k_i = t\left(\frac{1-\alpha}{2}, n_{sp}\,N_y - N_\theta\right)\sqrt{v_{ii}}\,, \quad i = 1,\dots,N_\theta, \quad (2.28)$$

where *t* is the upper $(1-\alpha)/2$ critical value for a *t*-distribution with $(n_{sp}\,N_y - N_\theta)$ degrees of freedom, n_{sp} is the number of experimental points, N_y is the number of model response and N_θ is the number of model parameters. For a $(1-\alpha) = 95\ \%$ confidence level, the confidence interval can be approximated by the following equation:

$$k_i^{95\%} = 2\sqrt{v_{ii}}\,, \quad i = 1,\dots,N_\theta. \quad (2.29)$$

For system without high correlation between model parameters, once the variance-covariance matrix of model parameters $\boldsymbol{V}_\theta$ is known, it is possible to carry out a t-test. The t-values are evaluated as:

$$t_i = \frac{\hat{\theta}_i}{\sqrt{v_{ii}}}, \quad i = 1,\dots,N_\theta, \quad (2.30)$$

where v_{ii} is the *i-th* diagonal element of $\boldsymbol{V}_\theta$. The *t-values* obtained from equation (2.30) are compared with respect to the reference *t-values* considering a *t-distribution* with $(n_{sp}\,N_y - N_\theta)$ degrees of freedom. High t-values usually mean that parameters are estimated with high confidence and high precision. When the system shows high correlation between model

parameters, it is necessary to take into account also the parameters covariance, thus not only the diagonal elements of matrix $\boldsymbol{V}_\theta$.
To verify the proper residual minimization, the *lack-of-fit-test* (χ^2-test) can be performed considering the sum of the weighted residuals:

$$SWR = \sum_{i=1}^{n_{sp}}\left[(y_i - \hat{y}_i)\Sigma_y^{-1}(y_i - \hat{y}_i)\right]. \quad (2.31)$$

In the *lack-of-fit-test* the SWR value is compared with a reference value from a χ^2 *distribution* considering $(n_{sp}\, N_y - N_\theta)$ degrees of freedom. If $SWR < \chi^{rif}$ the experimental data fitting is efficient and it is reasonable to consider the model as a reliable representation of the physical system. Because the assumption made in § 2.1.1 about the Gaussian distribution of measurements, the residual distribution should be confirmed by the "whiteness test".

2.2 Principal component analysis

Principal component analysis (PCA) is a mathematical and statistical technique used to formulate an empirical model derived from data that allows estimating one or more system properties from measurements.
The idea behind the PCA is to compress and then extract, in a proper way, the data available, making possible to (Wise and Gallagher, 1996):

- use only relevant data-embedded information, avoiding redundant information which are useless;
- describe how the system variables changes to each other's;
- averaging the data measurement, avoiding to describe the noise measurement.

Mathematically, the PCA relies on a decomposition of the original data set variance-covariance matrix. Let **X** the $m \times n$ matrix where the rows represents the samples while columns corresponds to variables. To avoid errors associated to different unit dimension between each variable, the data are pre-processed; each columns are mean-centered (i.e. the columns mean is subtracted to each column elements) and scaled to its variance (i.e. each column elements is divided by the columns variance).
The PCA decompose the data matrix X as:

$$\boldsymbol{X} = \boldsymbol{t}_1\boldsymbol{p}_1^T + \boldsymbol{t}_2\boldsymbol{p}_2^T + \ldots + \boldsymbol{t}_k\boldsymbol{p}_k^T + \boldsymbol{E} \quad (2.32)$$

where the $\boldsymbol{t}_i$ vectors are known as scores while $\boldsymbol{p}_i$ vectors are known as loadings. In equation (2.32), *k* is a number that necessarily must be less than or equal to the smaller dimension of **X**. The scores vectors contain the information of how the samples are related to each other, while loadings describe how the variables are related to each other.

For a given and auto-scaled matrix **X**, its covariance matrix can be defined as:

$$cov(\boldsymbol{X}) = \frac{\boldsymbol{X}^T\boldsymbol{X}}{m-1}\,. \tag{2.33}$$

The $\boldsymbol{p}_i$ vectors are the eigenvectors of the covariance matrix and by its definition:

$$cov(\boldsymbol{X})\boldsymbol{p}_i = \lambda_i\boldsymbol{p}_i\,, \tag{2.34}$$

where λ_i is the eigenvalue associated to its eigenvector. It is possible to think of λ_i as the amount of covariance described by the associated eigenvector. In the decomposition activity the $\boldsymbol{p}_i$ and $\boldsymbol{t}_i$ pairs are in descending order of related λ_i, so the first dimension describes the major variability. The *k* number of dimension used to describe the original data set can be chosen through the cumulative variance captured.
The scores vectors are obtained from the original data-set by the following equation:

$$\boldsymbol{t}_i = \boldsymbol{X}\boldsymbol{p}_i \tag{2.35}$$

The scores vectors $\boldsymbol{t}_i$ are orthogonal (i.e. $\boldsymbol{t}_i^T\boldsymbol{t}_j = 0\ for\ i \neq j$) to each other and form an orthogonal base, while the $\boldsymbol{p}_i$ vectors are orthonormal (i.e. $\boldsymbol{p}_i^T\boldsymbol{p}_j = 0\ for\ i \neq j$, $\boldsymbol{p}_i^T\boldsymbol{p}_j = 1\ for\ i = j$). Collecting the $\boldsymbol{t}_i$ vectors into the **T** matrix and the $\boldsymbol{p}_i$ vectors into the **P** matrix it is possible to represent the PCA decomposition as:

$$\boldsymbol{X} = \boldsymbol{T}\boldsymbol{P}^T + \boldsymbol{E}, \quad \boldsymbol{M} = \boldsymbol{T}\boldsymbol{P}^T\,, \tag{2.36}$$

where the product $\boldsymbol{T}\boldsymbol{P}^T$ represent the original data-set described by the new model and is function of the *k* number of selected dimension. Equations (2.32) and (2.36), thus the **X** matrix decomposition are shown in Figure 2.3 where the score and loading scores are highlighted.

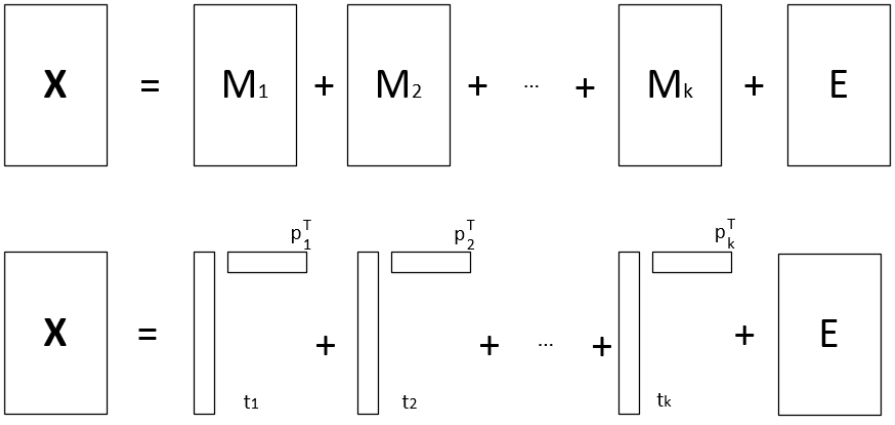


Figure 2.3 Representation of data matrix X decomposition as sum of M_i matrices and residual E. The M_i matrices are the products of the score and loading vectors product. (Adapted from Esbensen, Geladi, 2009).

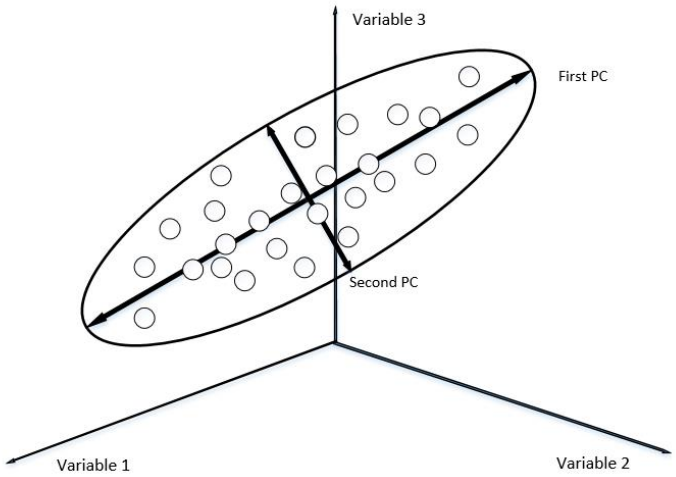


Figure 2.4 Representation of a projection of three dimensional space data in a two dimensional space detected by the two Principal Component.

The PCA permits a new base identification through scores vectors; it is possible to think of a new *k*-dimensional fictitious space where the original data-set is projected (Figure 2.4) (Geladi, 2009).

2.3 Supervised pattern recognition

The supervised pattern recognition problem can be represent as the wish of fit a model that properly relates the response to the observed predictors measurements (James *et al*, 2013) with the aim to accurately predicts the response for future observations. In contrast, in unsupervised pattern recognition problem there is no associated response to the observed measurements.
When quantitative variables are used as response ones, it is common refer to as a regression problem, while if the response variables are qualitative it is refer to as classification problems.
There are many classification techniques, also called classifiers, and all needs an observations training set which is used to build the classifier model. The classifier aim is to perform a correct classification not only in the training set, but also on the set not used during the training phase. To achieve this feature, the model is subject to the validation phase that can be done using an external data-set, or an internal data-set of the training set, which is referred to as cross-validation.

2.3.1 Support vector machine

In a *p*-dimensional space, let define the hyperplane as a flat affine subspace of *p*-1 dimension which mathematically is represented by the following equation (James *et al*, 2013):

$$\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p = 0 \qquad (2.37)$$

Given a $n \times p$ data matrix $\boldsymbol{X}$ representing n training observations of p-dimensional predictors

$$\boldsymbol{X} = \begin{pmatrix} x_{11} & \cdots & x_{n1} \\ \vdots & \ddots & \vdots \\ x_{1p} & \cdots & x_{np} \end{pmatrix}, \tag{2.38}$$

and assuming that each observation falls into two classes represented as $y_1, \ldots, y_n \epsilon\{-1, +1\}$, it is possible to construct a hyperplane on the form of equation (2.34) that perfectly separates the observations accordingly to their labels. Such hyperplane, for each i-observation has the following properties:

$$\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip} > 0 \;\; if \; y_i = 1, \tag{2.39}$$

$$\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip} < 0 \;\; if \; y_i = -1. \tag{2.40}$$

It is possible to think of the quantity on the right side of inequalities (2.36) and (2.37) as a measure of the distance of each point from the hyperplane.
The concept expressed by the equations (2.39) and (2.40) can be equivalently wrote as

$$\left(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip}\right) y_i > 0 \, . \tag{2.41}$$

Generally, if our data are perfectly separable, there will be exists an infinite number of hyperplanes that separates the two classes (Figure 2.5)

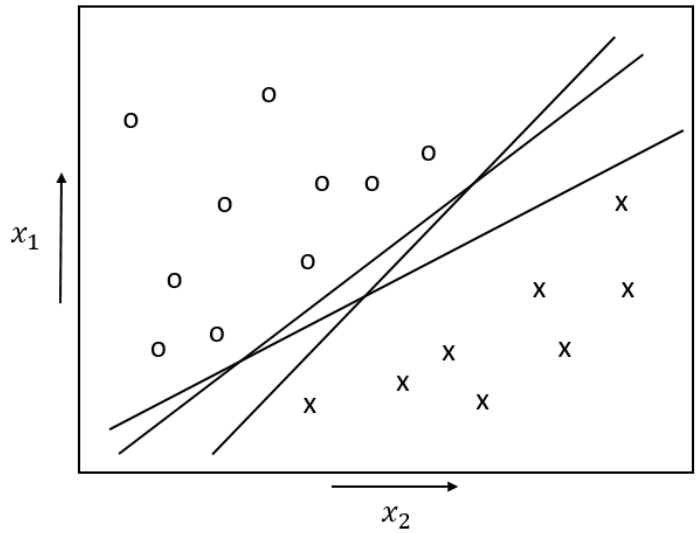


Figure 2.5 Three different hyperplane in a two-dimensional space that perfectly separates the observations belonging to two different classes, indicated respectively with circles and x.

Computing the perpendicular distance from each training observation to the hyperplane, and defining the margin as the minimum distance from observations to the hyperplane, the optimal separating plane is the one that maximize such distance. Such hyperplane is the solution to the following optimization problem:

$$\max_{\beta_0,\beta_1,\ldots,\beta_p} M, \tag{2.42}$$

$$\sum_{j=1}^{p} \beta_j^2 = 1, \tag{2.43}$$

$$\left(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip}\right) y_i > M \quad \forall\, i = 1, \ldots, n\,, \tag{2.44}$$

where M represent the margin to be maximized by changing the β_p hyperplane parameters. Equation (2.44) guarantees that each observation lies on the correct side of the hyperplane. Equation (2.43) impose that perpendicular distance from each observation to the hyperplane can be simply calculated by:

$$\left(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip}\right) y_i \tag{2.45}$$

Ensuring, together with equation (2.44) that each observation is on the correct side of the hyperplane and at least at a distance M from it.

The maximal margin hyperplane obtained from the equations (2.42), (2.43) and (2.44) resolution is function of only a small part of the training set. The observation that are fundamental for the problem resolution are called *support vectors* (Figure 2.6).

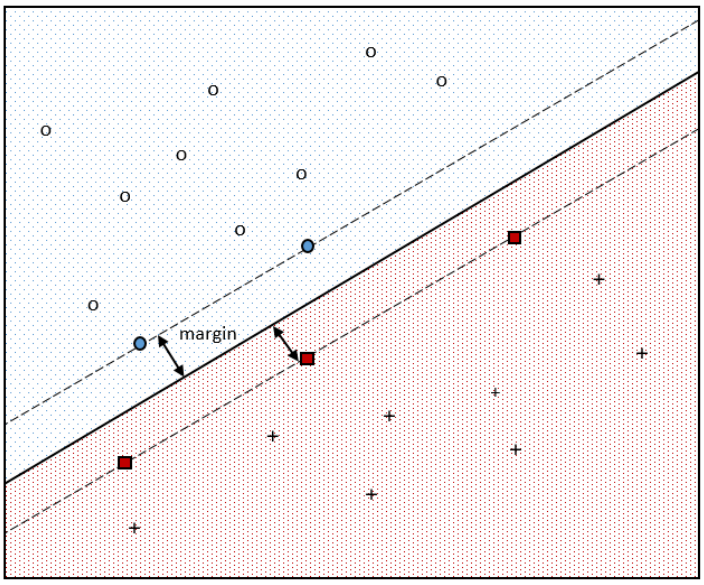


Figure 2.6 Maximal margin hyperplane representation (solid line) in a two-class problem (circles and pluses). The margin is the distance between the hyperplane and the dashed lines, which is detected by the support vectors shown as filled circles and filled square for each class, respectively.

The maximal margin hyperplane represented by equations (2.42), (2.43) and (2.44) is very sensitive to the observations and in some cases is desirable to have a classifier that does not perfectly separates the two classes in favour of the following features:

- greater robustness to individual observations;
- better classification of most of the training observations.

In order to obtain these features on the classifier method, the cost parameter (*C*) is added to the system through equation (2.44) that becomes:

$$\left(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip}\right) y_i > M\,(1 - \epsilon_i) \quad \forall\, i = 1, \ldots, n\,, \tag{2.46}$$

with the following constraints

$$\epsilon_i \geq 0,\ \sum_{i=1}^{n} \epsilon_i \leq C, \tag{2.47}$$

where the ϵ_i are slack variables that indicate where individual observations are located with respect to the hyperplane or the margin following this criterion:

- if $\epsilon_i = 0$ the i-th observation lies on the wright side of the margin;
- if $\epsilon_i > 0$ the i-th observation is on the wright side of hyperplane but violates the margin;
- if $\epsilon_i > 1$ the i-th observation is on the wrong side of hyperplane.

The *C* parameter bounds the sum of the slack variables and it represents the margin violations that can be tolerated; as *C* increase the margin violations will become more allowed, consequently the margin will be widen, in contrast low *C* values means wider and rarely violated margin (Figure 2.7).

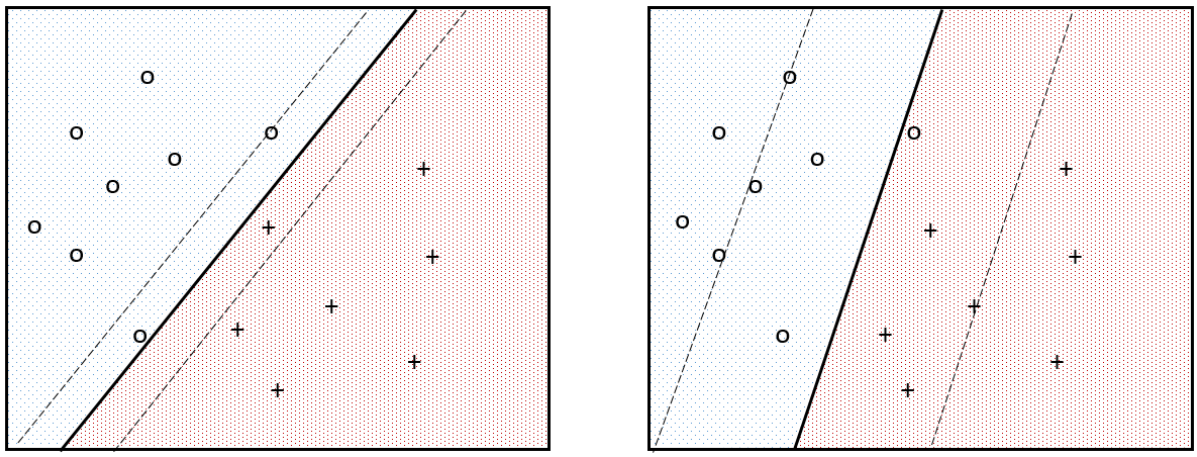


Figure 2.7 Representation of margin with respect to the cost paramenters. Low cost values produces wide margine (on the left). High cost values produces widen margin which are more violated (on the right).

2.3.2 Kernel functions

The introduction of kernel function in the maximal margin hyperplane problem permits to identify non-linear boundary between the two classes.

The solution of the optimization problem formulated by equations (2.42), (2.43), (2.46) and (2.47) involves the inner products of the observations vectors, which can be formulated as:

$$\langle x_i, x_j \rangle = \sum_{y=1}^{p} x_{iy}\, x_{jy}\ . \tag{2.48}$$

The linear support vector classifier can be represented as:

$$f(x) = \beta_0 + \sum_{i=1}^{n} \alpha_i \langle x, x_i \rangle, \quad (2.49)$$

where there are *n* parameters α_i, one for each training observation. As discussed earlier, only for the support vectors the α_i parameters are non-zero and this simplified the solution of the problem.

Generalization of the inner product can be written as $K(x_i, x_j)$, where *K* is some function called kernel, which quantifies the similarity between two observations.

Defining a kernel function in the support vector machine problem is possible to pass from the original input *p*-dimensional space to an enlarged dimensional space, where the samples are projected and the problem presented by equations (2.42), (2.43), (2.46) and (2.47) is solved (Figure 2.8) (Xu *et al*, 2006).

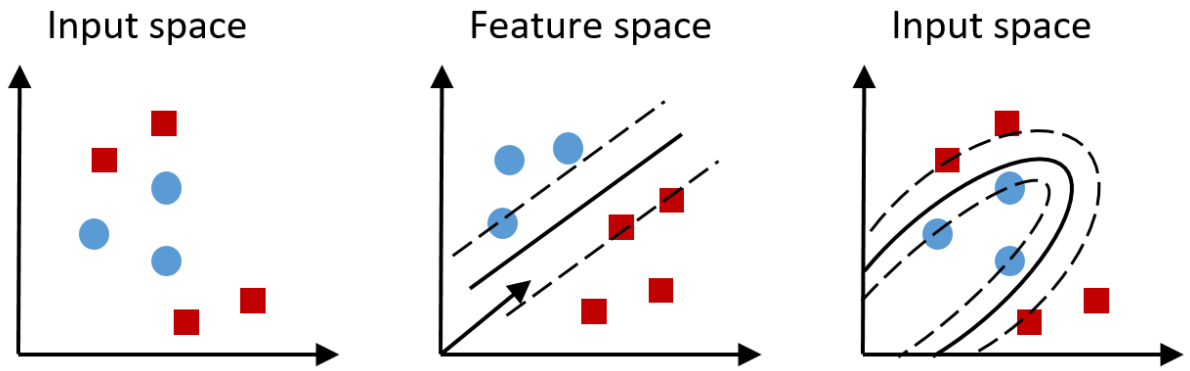


Figure 2.8 Representation of the support vector machine problem with non-linear kernel function. (Adapted from Xu, 2006).

Only certain kernels functions can be employed and are (Xu *et al*, 2006):

- Polynomial function (PF)

$$K(x_i, x_j) = (\alpha x_i^T x_j + b)^c \quad (2.50)$$

- Radial basis function (RBF)

$$K(x_i, x_j) = \exp \frac{-\|x_i - x_j\|^2}{2\sigma^2}; \quad (2.51)$$

- Sigmoidal function (SF)

$$K(x_i, x_j) = \tanh(\alpha x_i^T x_j + b). \quad (2.52)$$

Each kernel has a set of parameters that has to be tuned; the radial-basis-function, in particular, has a single parameter σ that represent the radial width which make it particularly widespread. Instead of use the σ parameter, it is common to use γ, which is related to σ following the relations:

$$\gamma = \frac{1}{2\sigma^2}\,. \tag{2.53}$$

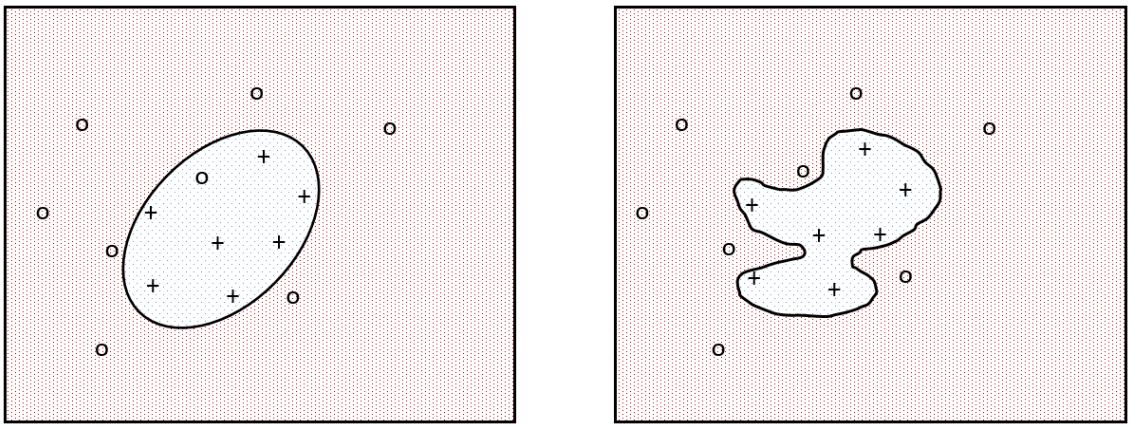


Figure 6.9 Representation of support vector machine problem boundaries with respect to the RBF kernel function parameter γ. Low values of γ gives smoother boundaries (on the left). High γ values gives more contorted boundaries (on the right).

The effects on boundaries with respects to the γ parameter are shown in Figure 2.9; as γ increases the boundaries complexity increases, too.

2.3.3 Model validation

Acting on the γ parameter, the kernel trick allows the definition of complex boundaries able to perfectly separates each observation belonging to different classes, but the model complexity must be controlled in favour of model robustness and also in order to avoid overfitting problems. The support vector machine optimization problem must be solved searching for a good compromise between the penalty error *C* and the kernel function parameters.
The solution is obtained through a cross-validation activity and the optimal set of parameters is the one that produce the lower misclassification error. Cross-validation is done splitting the *n* initial data set number of observations in *s* smaller datasets; some sets are used to train the model while the others to test its performance.
There are several type of cross validation, the most widespread are:

- venetian blinds, where the tests set is determined selecting every *s*-th observation in the original data-set;
- random subsets, where the test sets are determined randomly trough a selection of *n/s* observations, providing no presence of a single observation in more than one test set;
- leave-one-out, where single observation is used as test set.

The misclassification error is reported as a ratio between the number of misclassified subjects and the total subjects number classified. The misclassification error can be represented through

the confusion matrix, which is reported in Figure (2.10), where misclassification error is reported for each class analysed.

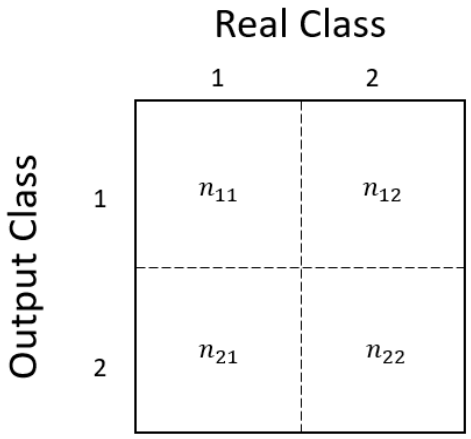


Figure 2.10 Confusion matrix representation for a two class classification problem.

In Figure 2.10, *n* is the number of subjects; the first subscript indicates the model predicted class, while the second subscript indicates the real subject class.

2.3.4 Probability estimation

There are two possible ways to assign an unknown observation to a class:

- based on the sign of the equation (2.39);
- based on the associated probability to belong of each class.

The probability distribution function is calculating through a parametric model fitting. Looking at the SVM problem resolution it is clear that some points (observation) are forced to be on the margin (support vector) and have all the same distance from the hyperplane. This fact involves a discontinuity in the class-conditional densities.
Because the class-conditional densities behaviour, the suggested probability distribution function (PDF) is the sigmoid form with two parameters on the form:

$$P(y = 1|f) = \frac{1}{1+exp(A\,f + B)}, \qquad (2.51)$$

where *A* and B are the sigmoid parameters.
The use of the entire SVM training set to train the sigmoid is to be avoided because it leads to a biased PDF model. The method used for deriving an unbiased training set is the cross-validation, which involves the training set splitting in three parts; in three-fold cross-validation, the training set is split into three parts and on two out of three parts the SVM model is trained,

while the remaining fold is used to evaluate the f_i, which is used to compute the probability distribution function expressed by equation (2.51). The union of all three sets of f_i form the entire sigmoid training set (Platt, 1999).

Chapter 3

Model-based type 1 VWD characterization

In this chapter, the VWD average subjects pharmacokinetic model parameters are compared to each other and discussed. The PK model parameters and PK indexes for type 1 VWD are compared to the healthy classes and the other VWD types and finally an intra-type 1 VWD categorization based on the mutation type are analysed through the PK model indexes.

3.1 Preliminary average VWD type 1 subject analysis

A critical point in the parameter estimation activity is the choice of initial guess for the parameters set. For this purpose, a fictitious VWD type 1 average subject has been created averaging the VWF:Ag and VWF:CB values at each sampling time for the entire VWD type 1 data-set available. The VWD type 1 data set has been given from the collaboration with the hospital of Padua.

The average VWD subject VWF:Ag and VWF:CB profiles for each VWD type is reported in Figure (3.1a) and Figure (3.1b). The fictitious VWF measurements, accordingly to Galvanin *et al* (2014), are assumed with a standard deviation $\sigma^{Ag} = \sigma^{CB} = 2\ IU/dL$.

Looking at the average VWF:Ag levels for the different VWD types (Figure 3.1a), the type 1 VWD subject is similar to the 2A and 2B average subjects and, with respect the healthy type, is more similar to the healthy O blood group type. The VWD type 1 average subject shows VWF:CB levels that are more similar to the healthy class O blood group than to those of the VWD type 2A and 2B.

Since the average heathy O blood group subject is the more similar to the VWD type 1 average subject, its set of PK model parameters has been used as initial guess to the average type 1 VWD subject parameters estimation. Subsequently the average VWD type 1 parameters set will be used as initial guess for the parameters estimation activity for each VWD type 1 subject.

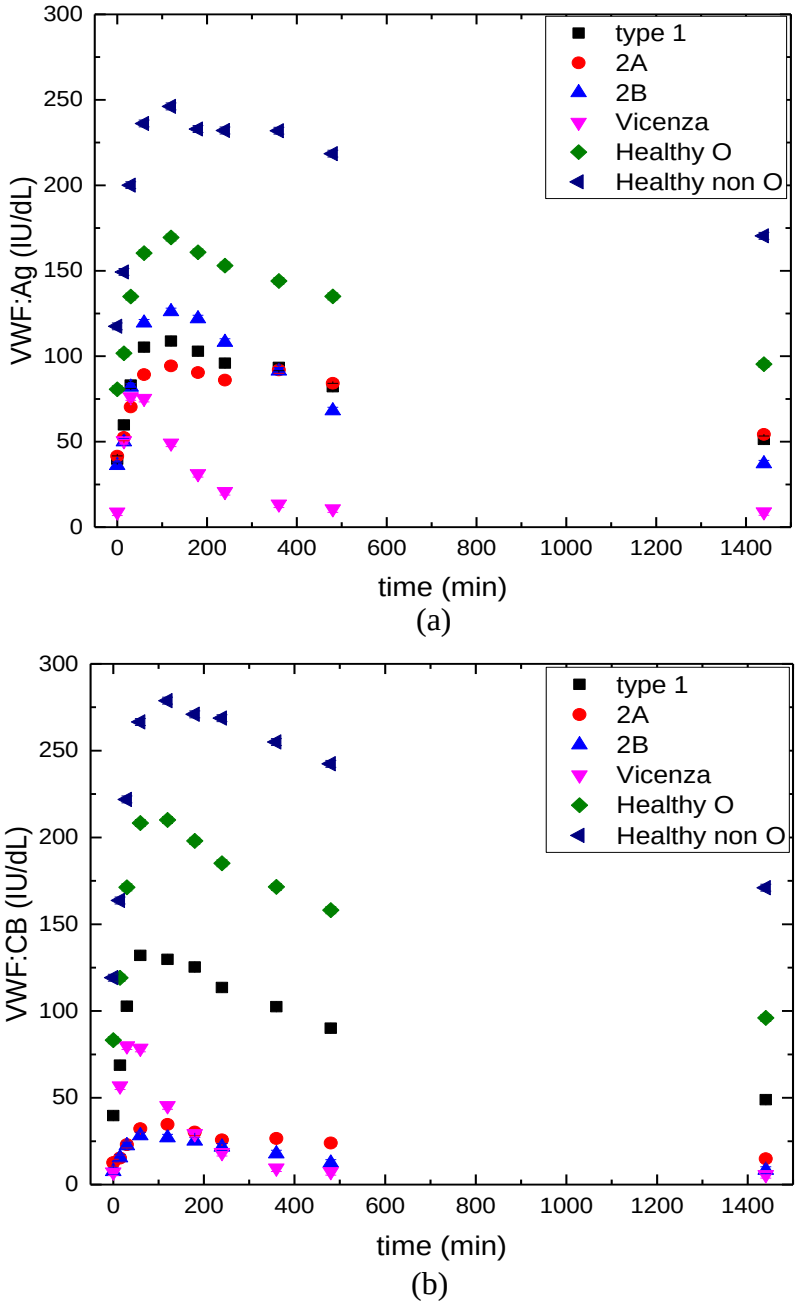


Figure 3.1 a) VWF:Ag levels for each average VWD categories and healthy classes. b) VWF:CB levels for each average VWD categories and healthy classes.

The parameters estimation of the pharmacokinetic model presented in § 2.1 is carried out using the gPROMS® v4.1.0 software. To facilitate the convergence the parameters estimation activity is split in three steps:

- *step 0*: all seven parameters $k_0, k_1, k_e, D, k, y_b^{CB}, D/t_{max}$ are free to vary;
- *step 1*: the corrective parameters k, y_b^{CB} and the D/t_{max} parameter are set at the values estimated in the previous step, while the kinetic parameters k_0, k_1, k_e, D are free to vary;

- *step 2*: the kinetic parameters k_0, k_1, k_e, D are set at the values estimated in the previous step, while the corrective parameters $k\ and\ y_b^{CB}$ are left free to vary.

Step 1 and *step 2* are repeated until the estimated values do not change significantly. The D/t_{max} parameter is kept constant for all the steps subsequently to step 0 since it is easily identifiable; this simplifies the parameter estimation process.

Because of the numerical issue that could arise due to the different scale associated of each parameter, the parameters estimation has been carried out considering a normalized set of parameters defined by the following equation:

$$\Theta_i = \frac{\hat{\theta}_i}{\mu_i}\ , \quad i = 1, \dots, N_\theta\ , \tag{3.1}$$

where μ is the parameter initial guess value, $\hat{\theta}$ is the original parameter and Θ is the normalized parameter. When the parameters are normalized, their sensitivities analysis measure the relative response variations, which are easily comparable to each other.

The parameter set for each average VWD type subjects are reported in Table 3.1. The type 1 VWD average subject shows a parameter set quite similar to the average healthy with O blood group, except for the release parameter *D* which is lower. Vicenza and 2B shows a higher proteolysis rate than the average VWD type 1.

Table 3.1 Pharmacokinetic models parameters set for each average VWD type subject (Castaldello, 2016).

Parameter	HnonO	HO	2B	Vic	Type 1
$\boldsymbol{k_0}$ $[\boldsymbol{h^{-1}}]$	0.0285	0.0264	0.0136	0.0423	0.0280
$\boldsymbol{k_1}$ $[\boldsymbol{h^{-1}}]$	0.00014	0.00033	0.0021	0.0013	0.0004
$\boldsymbol{k_e}$ $[\boldsymbol{h^{-1}}]$	0.00069	0.00136	0.00369	0.00983	0.0014
$\boldsymbol{D}$ $[IU]$	167.631	140.511	193.638	198.938	86.4874
$\boldsymbol{k}$ [-]	0.9681	0.9952	0.2231	0.7427	0.9418
$\boldsymbol{y_b^{CB}}$ $[\boldsymbol{IU/dL}]$	84.670	50.675	29.464	5.6121	33.0138
$\boldsymbol{D/t_{max}}$ *[IU/h]*	1.5436	1.2669	0.85196	2.9623	0.7582

Average VWD type subjects cannot be considered a good representation of the entire VWD type because of the high variability (Taverna, 2017), anyway the parameters set reported in Table 3.1 suggest that the PK model is capable to represent the different VWD types. The high Vicenza and 2B proteolysis rate means a fast reduction in high VWF multimers weight and a reduced VWF haemostatic ability. The low release parameter means a quantitative reduction in VWF, which characterized the VWD type 1.

The average VWD type 1 subject parameter estimation is satisfactory in terms of precision; in Table 3.2 and Table 3.3, respectively, the parameters estimation statistics are reported for *step 1* and *step 2*. All the parameters satisfy the t-test and the k_1 parameter shows the lower t-value with respect to the other parameter.

Table 3.2 Parameters estimation results for *step 1*. The *t-test* and *chi-squared* results are reported.

	Initial Guess	**Final Value**	**95% t-value**	**Dev.st.**	**Reference t-value (95%)**	**Weighted Residual**	χ^2 **(95%)**
k_0	1.0557	1.0557	11.69	0.04261			
k_1	1.08965	1.08965	2.639	0.1948	1.74561	58.4853	26.2962
k_e	1.05573	1.05573	8.024	0.06206			
D	0.612245	0.612245	13.91	0.02076			

Table 3.3 Parameters estimation results for *step 2*. The *t-test* and *chi-squared* results are reported.

	Initial Guess	**Final Value**	**95% t-value**	**Dev.st.**	**Reference t-value (95%)**	**Weighted Residual**	χ^2 **(95%)**
k	0.954804	0.954788	13.58	0.03333			
y_b^{CB}	0.660656	0.660641	8.584	0.03647	1.73975	58.4853	27.5871
D/t_{max}	0.592401	0.592373	115.4	0.002433			

The χ^2 test is not satisfied, in fact χ^2 < SWR. Looking at the predicted VWF:Ag and VWF:CB profiles with respect the experimental values (Figure 3.2), the model response well captures the general experimental behaviour.

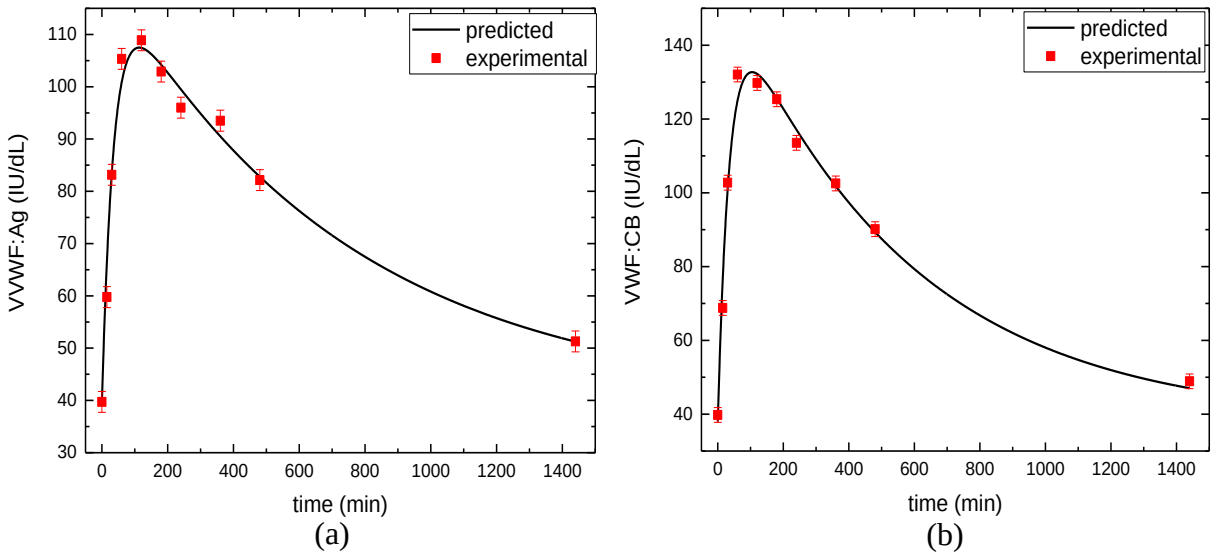


Figure 3.2 a) Experimental and model predicted VWF:Ag levels for the average VWD type 1 subject. b) Experimental and model predicted VWF:CB levels for the average VWD type 1 subject.

The sensitivity analysis (Figure 3.3a) shows that the k_1 and y_b^{CB} parameters do not affects VWF:Ag measurements; k_1 do not affect the VWF:Ag levels because of the model formulation while y_b^{CB} is related to VWF:CB measure only. The sensitivity analysis suggests that the k_1 and y_b^{CB}parameters cannot be identified from the VWF:Ag measures but they are exclusively related to the VWF:CB measures. Looking at the sensitivity analysis with respect to the VWF:CB model response (Figure 3.3b), the k_1 parameter may exhibit some identifiability issues since it is not significantly affected by the VWF:CB model response.

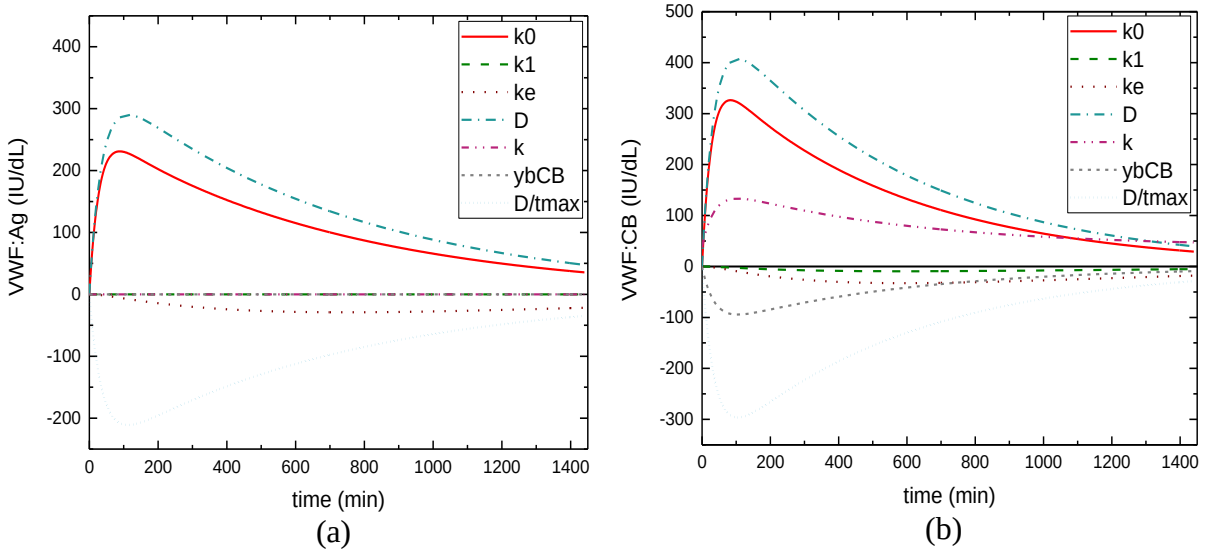


Figure 3.3 a) Representation of VWF:Ag model response sensitivity to the entire parameters set. b) Representation of VWF:CB model response sensitivity to the entire parameters set.

A heteroscedastic variance model presented in §2.11 has been used to quantify the VWD type 1 group variability. The models parameters are obtained by the minimization of the sum of squared residual and the results are reported in Table 3.4 for each VWD category.

Table 3.4 Heteroscedastic variance model parameters.

Response	Parameter	H0	Hnon0	2B	Vic	Type 1
VWF:Ag	ω	7.917	20.589	4.418	1.252	1.180
	γ	0.325	0.210	0.397	0.762	0.7893
VWF:CB	ω	2.436	17.090	0.128	2.028	0.9680
	γ	0.578	0.258	1.533	0.653	0.8477

At the maximum VWF:AG and VWF:CB levels for average type 1 subject, which are respectively about 110 IU/dL and 135 IU/dL, the model standard deviation is respectively 58 IU/dL and 62 IU/dL.

3.2 Model based VWD type 1 characterization

Using the average VWD type 1 subject parameters reported in §3.1 as initial guess, the model parameter estimation for each VWD type 1 subject has been carried out. The estimated PK parameters are averaged and compared with respect the averaged parameters for each VWD type. The healthy classes, which are divided respectively in O and non-O blood group, are used as control groups. The group variance are compared with the Tukey-Kramer test assuming a studentized range distribution. The null hypothesis is that the two groups analysed can be considered different. The null hypothesis is rejected if:

$$|t| = \frac{|y_i - y_j|}{\sqrt{MSE\left(\frac{1}{n_i}+\frac{1}{n_j}\right)}} > \frac{1}{\sqrt{2}}\, q_{a,k,N-k}, \tag{3.2}$$

Where $q_{a,k,N-k}$ is the upper 100*(1-α)*th* percentile of a studentized range distribution, *k* is the number of groups and *N* is the total number of observations. The value *N*-*k* represents the degrees of freedom. The chosen *p*-value used in the Tukey-Kramer test is 0.05.

VWD type 1 shows an average elimination rate higher than both the healthy classes even if statistically can not be considered different from healthy O blood group class while is different from healthy non-O blood group class (Figure 3.4a). Vicenza VWD type shows the highest average elimination rate k_e while VWD type 2B has a reduced average k_e with respect the Vicenza but higher than the healthy classes and type 1 VWD. Both Vicenza and 2B type are statistically different from healthy and type 1 classes.

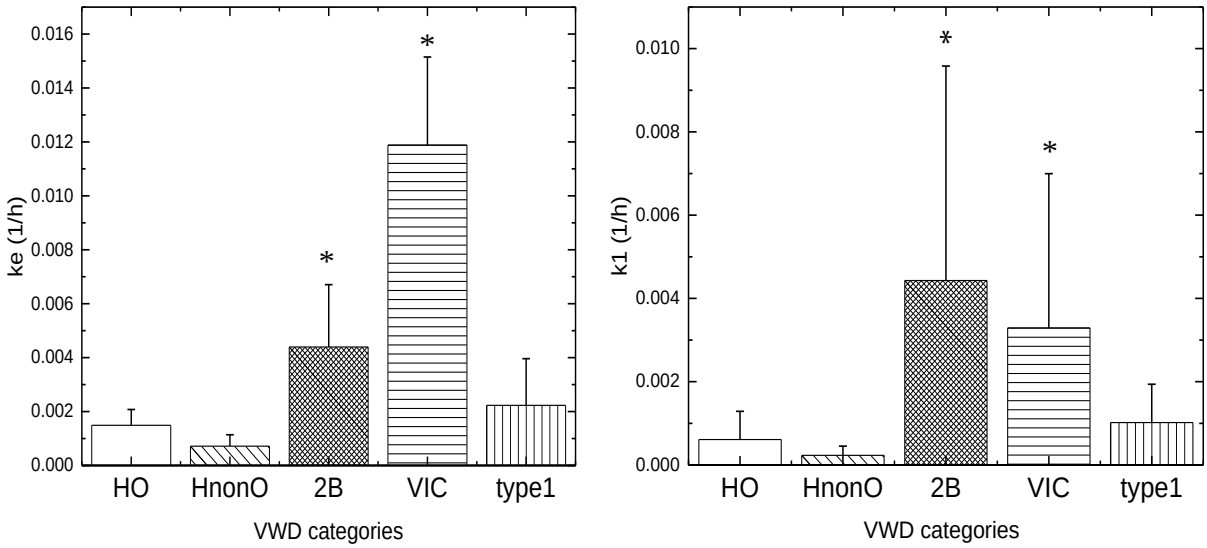


Figure 3.4 a) Comparison between the estimate elimination rate for each VWD type and healthy class. b) Comparison between the estimate proteolysis rate for each VWD type and healthy class. Asterisks indicate the parameters significantly different from control groups ($p < 0.005$).

High elimination rate means a fast reduction in both UL+H and L PK model compartments. A fast UL+H VWF multimers weight elimination involves a reduction in the VWF haemostatically action which characterized the Vicenza and 2B VWD type.

Type 2B and Vicenza shows an enhance proteolysis rate k_1 (Figure 3.4b), which means a fast reduction in the high VWF multimers weight; even if the standard deviation is high, the k_1 for these classes can be considered statistically different from both the healthy and type 1 classes.

VWD type 1 has an average k_1 comparable to the healthy classes while, for the healthy classes, subjects with non O blood group shows a reduce average proteolysis rate. Statistically the healthy and type 1 VWD classes cannot be considered independent. The results shown in Figure 3.4b do not consider some subjects affected by VWD type 1 because of the difficulties related to the k_1 identification.

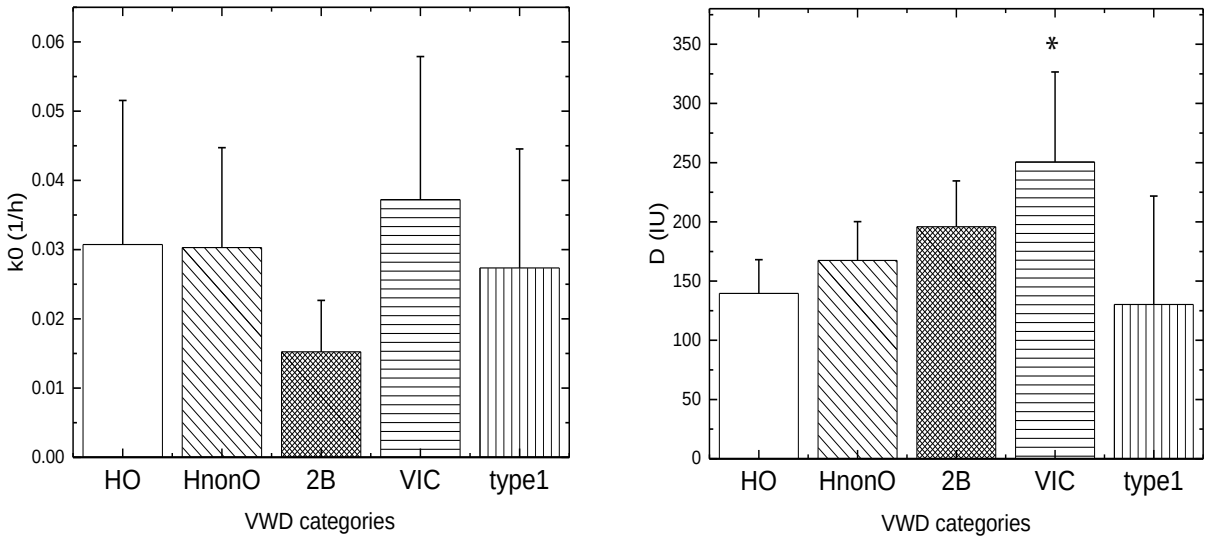


Figure 3.5 a) Comparison between the estimate release rate for each VWD type and healthy class. b) Comparison between the estimate release parameter for each VWD type and healthy class. Asterisks indicate the parameters significantly different from control groups ($p < 0.005$).

Average k_0 for all the VWD type do not differs significantly, except for the 2B type which shows a lower average release rate. By the way, statistically none of the classes can be considered different from each other (Figure 3.5a).

The average release parameter *D* (Figure 3.5b) for VWD type 1 is lower than both the healthy classes but its high variability make it not statistically different from healthy classes. The 2B and Vicenza types show high *D* parameter, but only the subjects affected by VWD type Vicenza can be considered statistically different from the healthy classes. k_0 and *D* are two parameters that identify the VWF release which can not be measured and their effect on the phenomena representation is not always clearly distinguishable; the amount of VWF released, instead, takes into account both the parameters and the combined effects. The average amount of VWD released for VWD type 1 is slightly reduced with respect the healthy class O blood group which is lower than the healthy class non O blood group (Figure 3.6a). The 2B type has the highest average Q_{re} , while the Vicenza group is slightly higher than the VWD type 1. Statistically, only VWD type 2B can be considered different from the healthy classes, while the type 1 VWD cannot be considered different from type Vicenza. This result suggests that the Vicenza type could be classified under VWD type 1 (N.H.B.L.I., 2007).

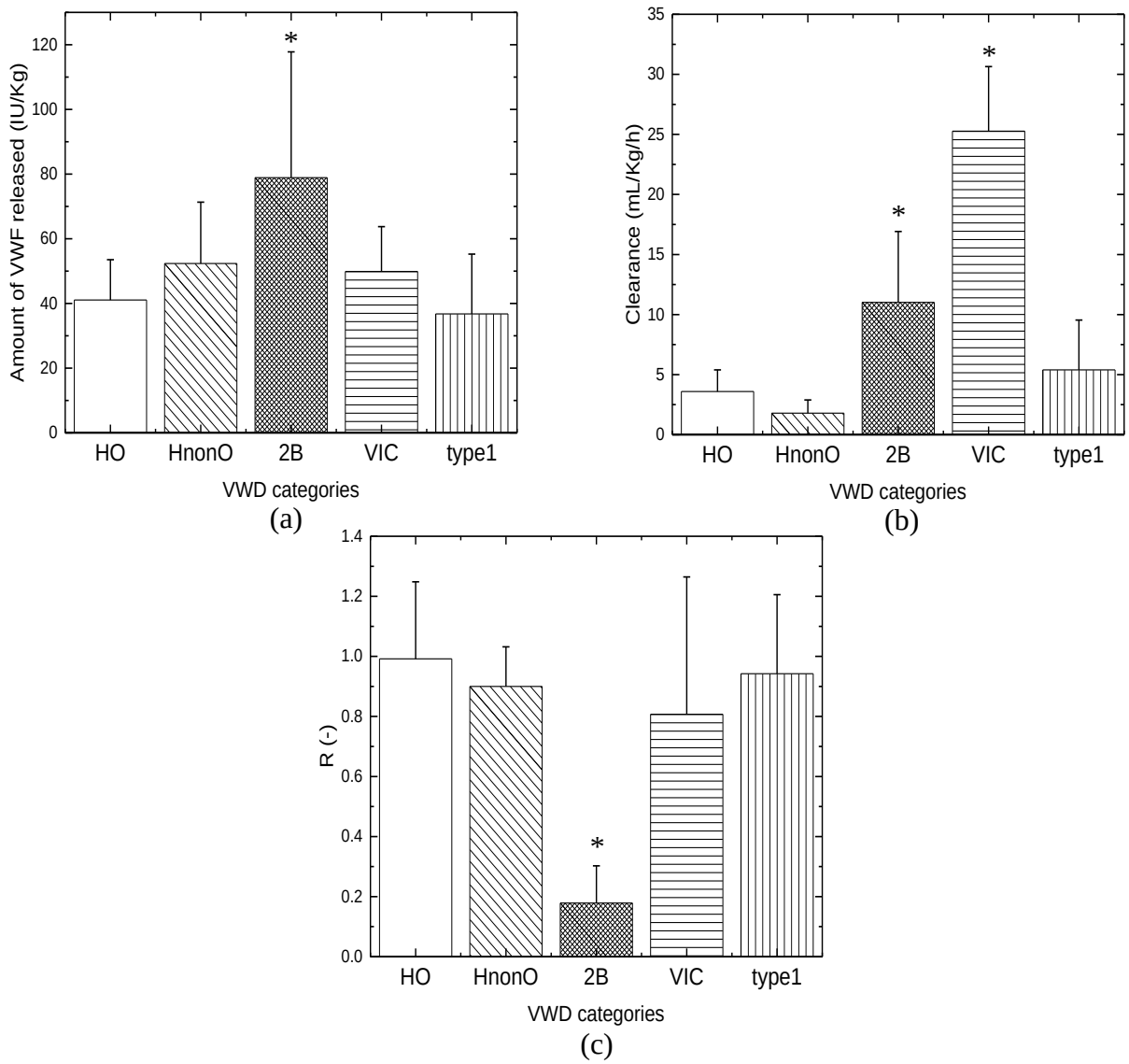


Figure 3.6 a) Comparison between the amounts of VWF released for each VWD type and healthy class. b) Comparison between the clearance for each VWD type and healthy class. c) Comparison between the ratio for each VWD type and healthy class. Asterisks indicate the parameters significantly different from control groups ($p < 0.005$).

VWD type 1 shows an average clearance higher than both the healthy classes; the healthy O blood group has an enhanced clearance with respect the non-O. Statistically the VWD type 1 category can be considered different from healthy non-O group but not from healthy O group. Both the 2B and Vicenza type have a high clearance with respect the healthy classes (Figure 3.6b).

Looking at the ratio (*R*) at the basal condition between the VWF:CB and VWF:Ag levels (Figure 3.6c) the VWD type 1 shows an average value similar to the other classes except for type 2B which shows the lowest average ratio.

The model-based results, in terms of amount of VWF released and clearance, confirm the knowledge found in literature (Sadler, 2014). VWD type 1 shows a modest decrease amount of

VWF and an increased clearance with respect to the healthy subjects. With respect the other VWD types, type 1 is the most similar to the healthy class with O blood group, confirming the difficulties related to its detection and correct diagnosis.
In Figure 3.7 the combination of amount of VWF released and clearance is reported for each healthy and affected by VWD subject classified for VWD categories. All the VWD Vicenza subjects shows a higher clearance than the others categories. Some subjects of type 2B VWD shows reduced clearance, which could be considered similar to some type 1 VWD and healthy O blood group subjects. Healthy subjects are well characterized based on the blood group; non-O blood group subjects shows a lower value of clearance with than healthy O blood group.

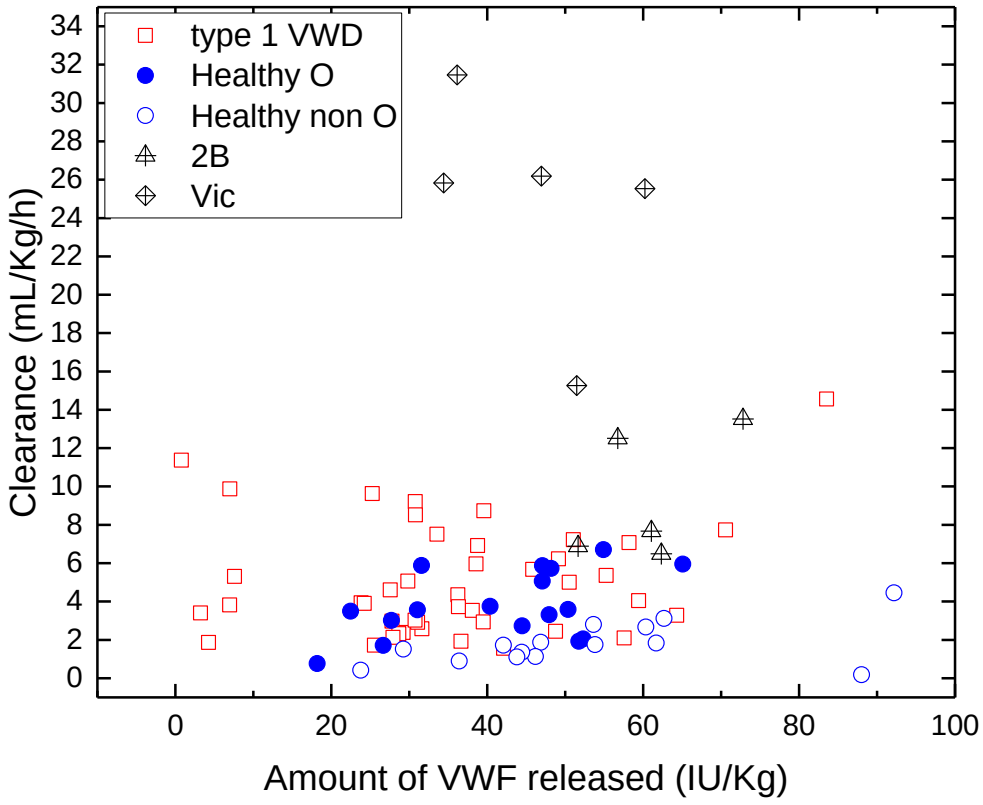


Figure 3.7 Combined model-based pharmacokinetic parameters amount of VWF released and clearance for all the analysed subjects divided by the VWD types and healthy classes.

The region that identify the type 1 VWD is quite overlapped with the healthy classes; subjects with non-O blood group are ore only partially overlapped, while O blood group subjects are not clearly distinguishable basing on the proposed PK indexes.

3.3 Intra VWD type 1 characterization

The subjects affected by VWD type 1 are divided into categories depending on the mutation type in the VWF gene. The data has been collected for the three categories: missense mutation, nonsense mutation and no mutation.

The elimination rate k_e is increased for all the VWD type 1 categories (Figure 3.8a). Healthy O group shows an increased clearance with respect the heathy non O blood group. The subjects without mutation shows the highest average clearance and it is the only category that can be considered statistically different from both the healthy classes. The missense and nonsense group statistically differs from healthy non O group but not from healthy O group.

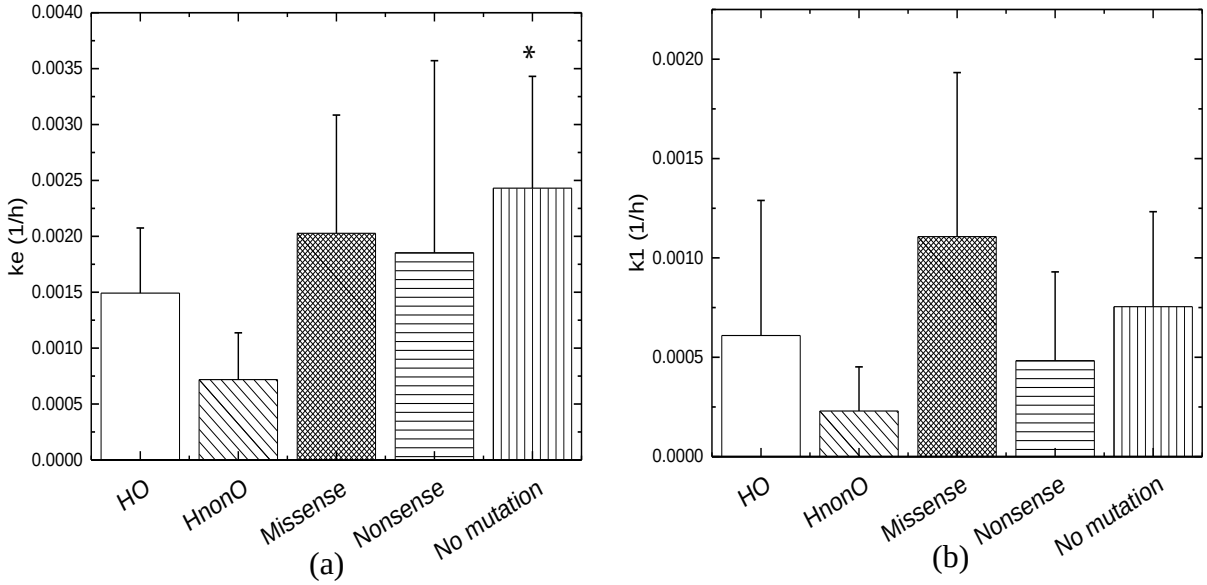


Figure 3.8 a) Comparison between the estimate elimination rate for each type 1 VWD mutation categories and healthy class. b) Comparison between the estimate proteolysis rate for each type 1 VWD mutation categories and healthy class. Asterisks indicate the parameters significantly different from control groups ($p < 0.005$).

Looking at the average k_1 for each class (Figure 3.8b), the missense mutation group shows the highest value; because of its high variability it can be considered statistically different from healthy non-O blood group but not from the healthy O blood group. Healthy O group, in fact, shows a higher average proteolysis rate than the healthy non-O blood group. Nonsense mutation and no mutation categories have an average k_1 similar to the healthy O blood group but statistically are not different. The results shown in Figure 3.7b did not take into account the entire VWD type 1 dataset because of the difficulties related to the proteolysis rate estimation. Looking at both average k_0 and average D for each VWD type 1 categories (Figure 3.9), there is no appreciable differences between healthy and VWD type 1 classes. Furthermore, none of the VWD categories can be considered statistically different from the healthy classes.

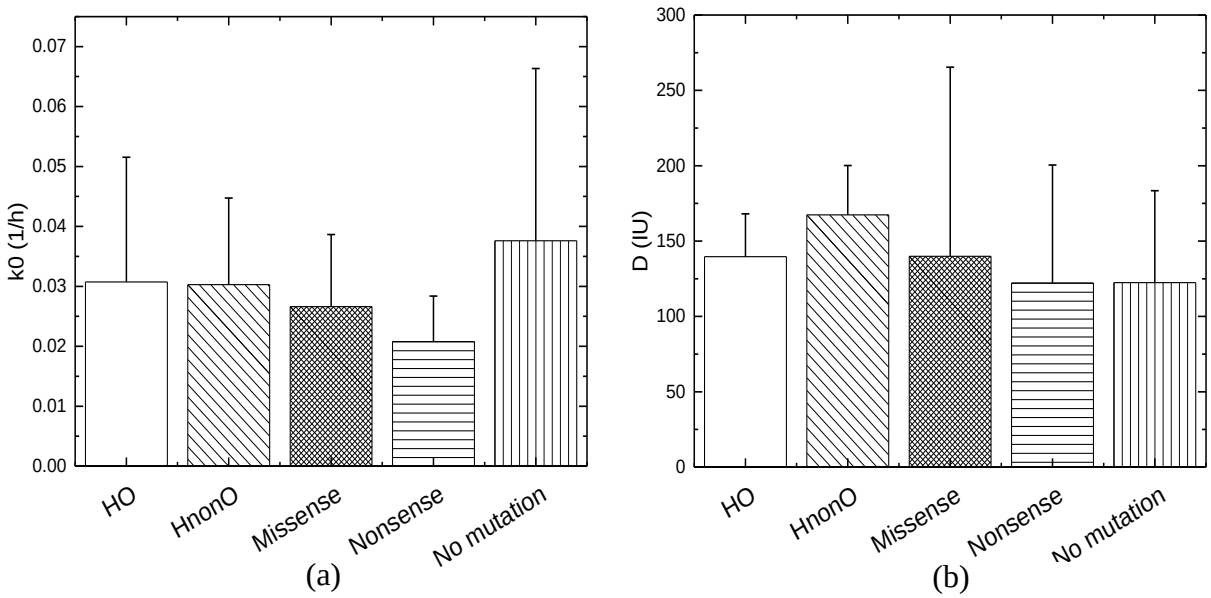


Figure 3.9 a) Comparison between the estimate release rate for each type 1 VWD mutation categories and healthy class. b) Comparison between the estimate release parameter for each type 1 VWD mutation categories and healthy class.

The average amount of VWF released for the missense and nonsense mutation groups is lower than the healthy groups (Figure 3.10a). Subjects with no VWF gene mutation shows an average VWF release, which is slightly higher than the healthy O class and lower than the healthy non-O class. Looking at the variability intra-group, none of the VWD type 1 categories can be considered different from the healthy O blood group, while the two mutation groups are statistically different from healthy non-O blood group (Figure 3.10a).

The subjects with VWF gene mutation shows a higher clearance than the healthy classes but slightly lower than the subjects without mutations; because of their variability the VWD type 1 categories cannot be considered statistically different from each other. Furthermore, none of the VWD type 1 categories can be considered statistically different from healthy O class. The no mutation category shows the highest average clearance and is the only VWD type 1 category statistically different from the healthy non-O blood group class (Figure 3.10b).

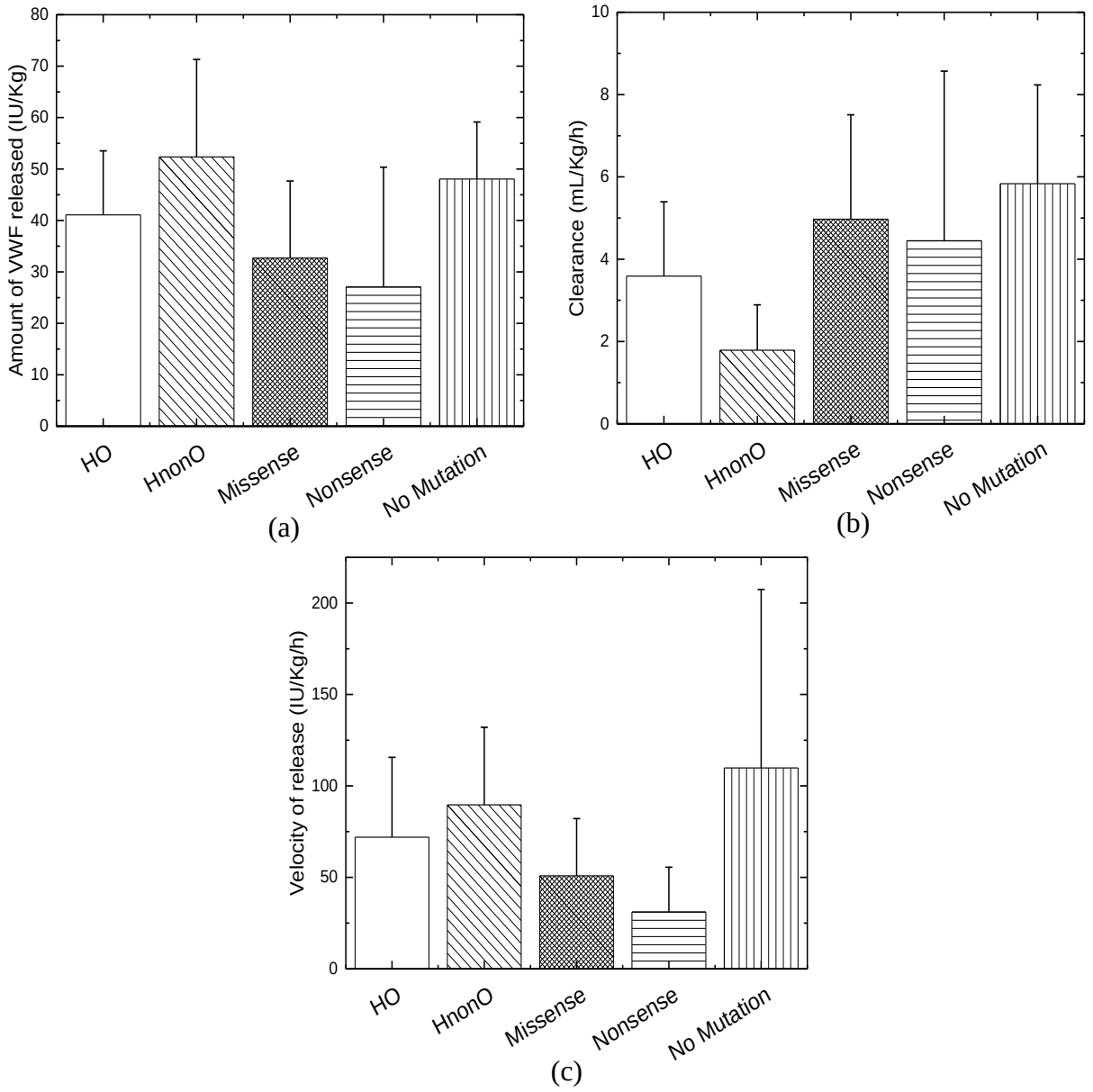


Figure 3.10 a) Comparison between the amounts of VWF released for each VWD type 1 categories and healthy classes. b) Comparison between the clearance for each VWD type 1 categories and healthy classes. c) Comparison between the ratio for each VWD type 1 categories and healthy classes.

The subjects without mutation shows the highest average velocity of release which can be considered statistically different from subjects with gene mutation. The missense and nonsense categories shows an average velocity of release lower than the healthy classes, even if it cannot be considered statistically different because of the high variability (Figure 3.10c).

Looking at Figure 3.11 where the combination of the amount of VWF released and the clearance is reported for each VWD type 1 subject distinguished by genetically categorization, it is possible to recognise three main clusters:

1. a cluster characterized by a very low amount of VWF released (<15 IU/kg) (red ellipse) comprising subjects 6, 8, 17, 19, 33 and 48. In the literature this cluster is referred to as severe type 1 (Sadler, 2003), (Casonato *et al*, 2016);
2. a cluster characterized by a slightly reduced amount of VWF released and a reasonable increased clearance (black ellipse), which comprise mainly subjects carrying missense and nonsense mutation and subject 27, which does not carry any mutation;
3. a cluster characterized by an increased amount of VWF released and increased clearance which (blu ellipse) mainly include subjects without mutation and some subjects carrying missense mutations.

Cluster 2 and cluster 3 together identify the mild type 1 category.

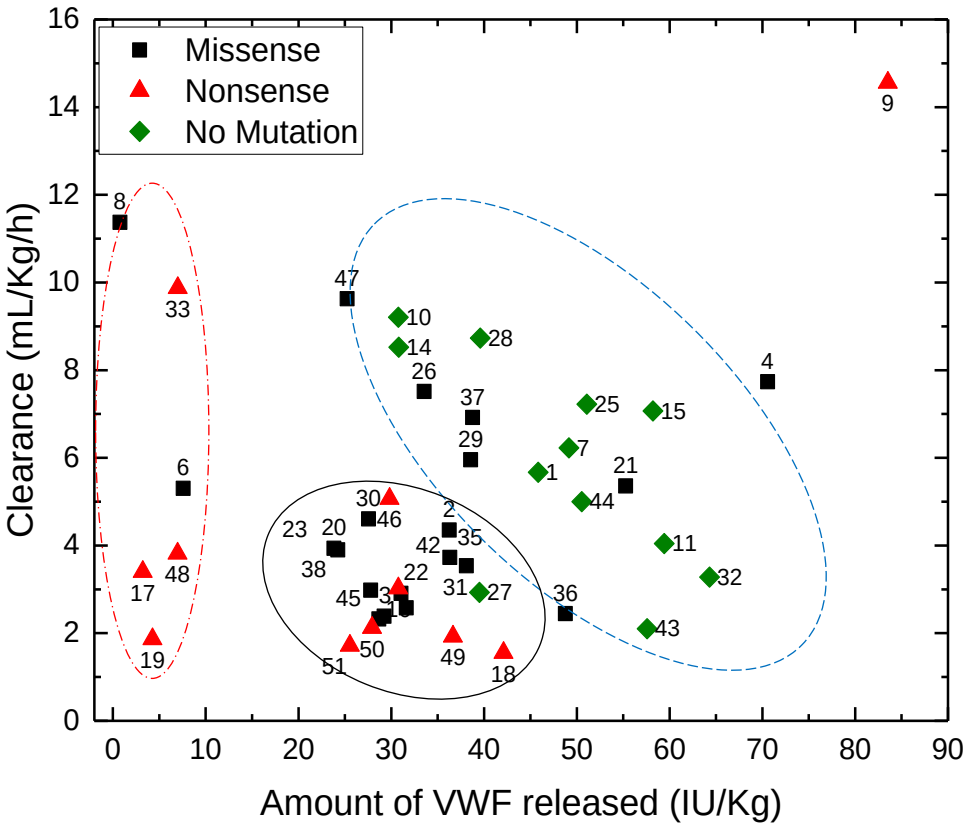


Figure 3.11 Combined model-based pharmacokinetic parameters amount of VWF released and clearance for the entire type 1 VWD set of subjects. The three main clusters are highlighted.

The high variability associated to each VWD type 1 categories is clearly visible in Figure 3.11; the subjects that carry nonsense mutation are present in both mild and severe VWD type 1

categories depending on the specific mutation. Subjects 17 and 19 carry a double VWF gene mutation while subject 48 carry a nonsense homozygote mutation.

The subjects carrying missense mutation are mainly in the mild type 1 clusters except subject 6 which shows a gene deletion and subject 8. Some subjects carrying missense mutation are quite similar in terms of amount of VWF release and clearance to subjects without VWF gene mutation; subjects 47 and 29 carry the C1130F mutation and shows an enhanced clearance. The high clearance shows by the PK model is confirmed by the literature (Sadler, 2014). Subjects 4, 26 and 36 carry the P2063S missense VWF gene mutation but its correlation with the VWD is not clear (Hampshire, 2013); so it is reasonable to think of this subjects more similar to the subjects without mutation.

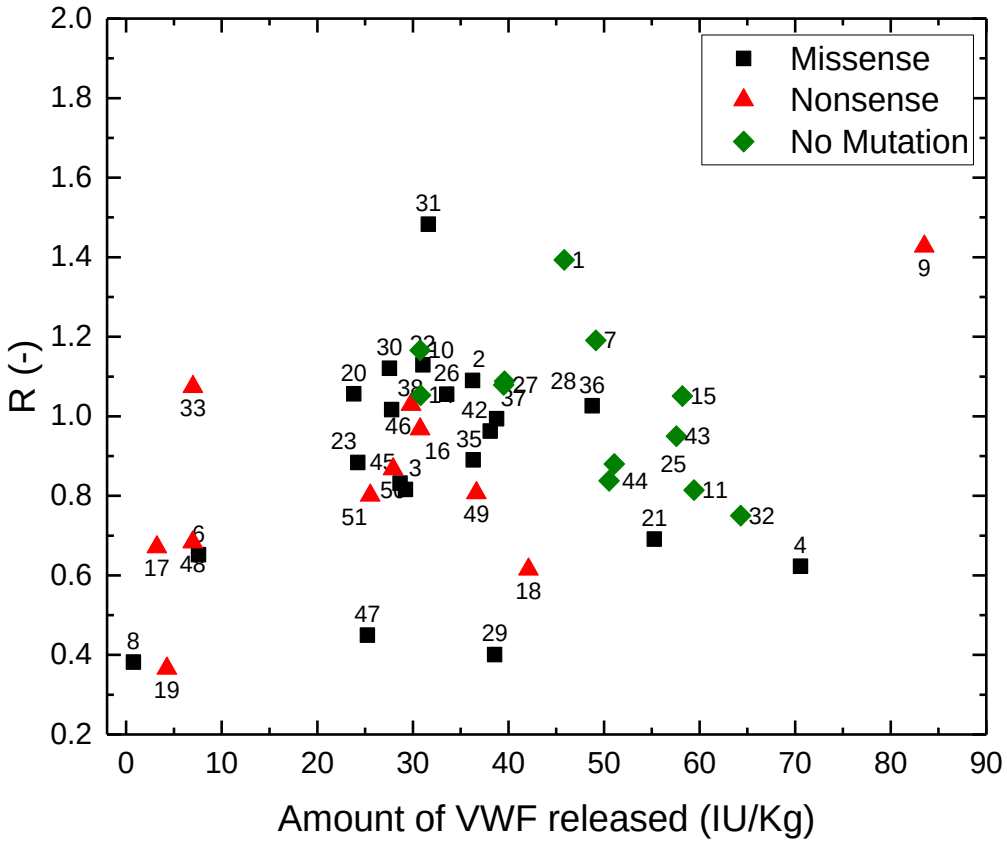


Figure 3.12 Combined model-based pharmacokinetic parameters amount of VWF released and VWF:CB and VWF:Ag ratio at the basal state for the entire type 1 VWD set of subjects.

Looking at the combined measures of amount of VWF released and the basal state ratio between VWF:CB and VWF:Ag levels (Figure 3.12) it is possible to recognise the same three clusters identify by the combination of amount of VWF released and clearance. An important overlapping between the mild type 1 clusters is present; subjects 10, 14 and 27, which did not carry any mutation shows an amount of VWF released and a ratio similar to the missense and nonsense mutation subjects. Subjects 29 and 47 are the only subjects present in the data set,

which bring the C1130F VWF gene mutation. They are comprised in the mild type 1 VWD but show a reduced basal ratio suggesting a partial reduction in the VWF high multimers weight. Figure 3.13 shows how type 1 VWD subjects and healthy subjects are located with respect to the clearance and amount of VWF released. A partial overlapping between the region identified by type 1 subjects and the one identified by healthy non-O blood group subjects is present; subjects 32 and 43 do not carry any VWF gene mutation, while subject 18 carries a nonsense mutation.

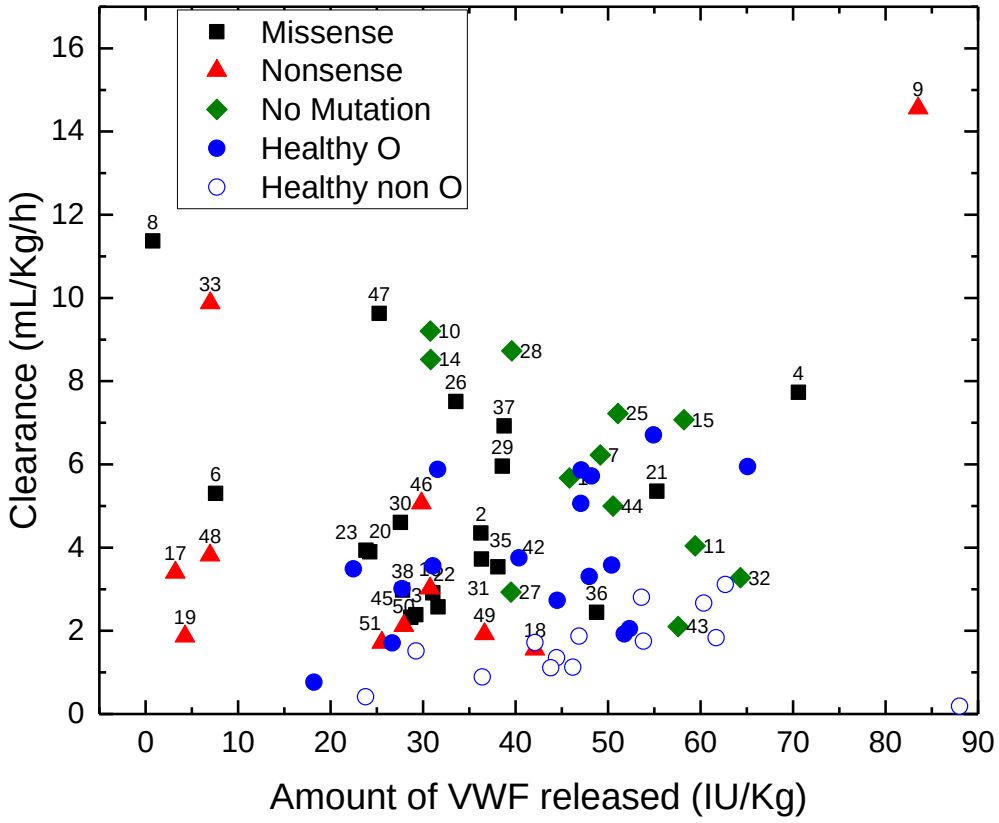


Figure 3.13 Combined model-based pharmacokinetic parameters amount of VWF released clearance for the entire type 1 VWD and healthy set of subjects.

Subjects with severe type 1 VWD shows an amount of VWF released lower than the healthy subjects and subjects 8 and 33 also exhibit a higher clearance. The subjects with VWF gene mutation belonging to the mild type 1 VWD show quite the same PK indexes as the healthy O blood group subjects. Looking at the subjects without VWF gene mutation, a partial overlapping with both healthy O e non-O blood groups is present.

Chapter 4

Model-based type 1 VWD diagnosis

In this chapter, the proposed algorithm to obtain the type 1 VWD diagnosis is presented. The data and the data analysis used in the algorithm are described and each algorithm step is validated through a cross-validation method; the results in terms of performances are reported and discussed.

4.1 Preliminary raw experimental data analysis

In Chapter 3 the pharmacokinetic model results demonstrated the PK model difficulty to describe the differences between healthy and affected subjects. A PCA was carried out on the entire autoscaled dataset of experimental VWF:Ag and VWF:CB levels during the DDAVP test considering all the VWD types and the healthy classes. The variability captured by the firsts two PC for the entire dataset is reported in Figure 4.1. The cumulative variance captured is the 93.11% of the total.

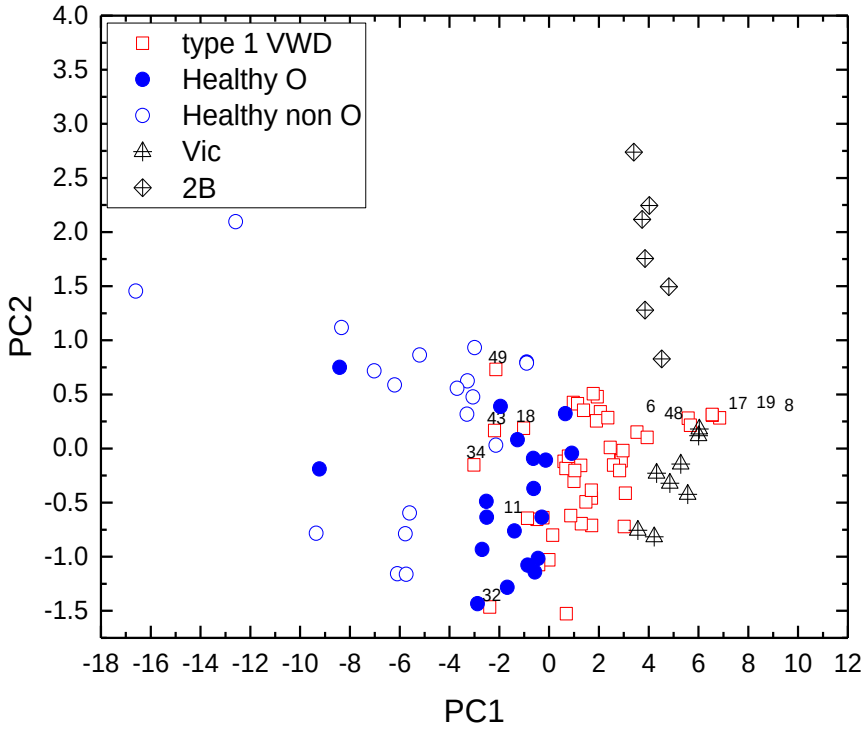


Figure 4.1 Representation of the firsts two principal component of a PCA carried out on the experimental VWF:Ag and VWF:CB measures considering all the VWD subtypes and healthy subjects.

Type 1 VWD subjects show VWF:Ag and VWF:CB levels between the healthy O blood group and Vicenza VWD type, while the 2B group is quite separated from all the other categories. Looking at the raw experimental VWF:Ag and VWF:CB measures, subjects 6, 8, 17, 19 and 48, which are affected by severe type 1 VWD, show VWF levels that are similar to the Vicenza type while the model-based PK indexes shown in previous chapter seems able to better capture the differences between the two categories in terms of clearance and *R*.

Focusing only for the type 1 VWD affected subjects, a PCA is carried out only on the raw experimental data belonging to type 1 VWD and the two healthy classes, and the variability captured by the firsts two PC is reported in Figure 4.2. The subjects belonging to the healthy class are 38, while the subjects affected by type 1 VWD are 51. The cumulative variance captured is the 93.88% of the total.

A partial overlapping is present between type 1 VWD subjects and healthy O blood group subjects. The subjects with VWF:Ag and VWF:CB levels more similar to the healthy classes are mainly subjects with P2062S VWF gene mutation. Subjects 2, 4, 26 and 36 carry the P2063S mutation, subject 43 do not carry any VWF gene mutation while subjects 18 and 49 both carry nonsense mutation.

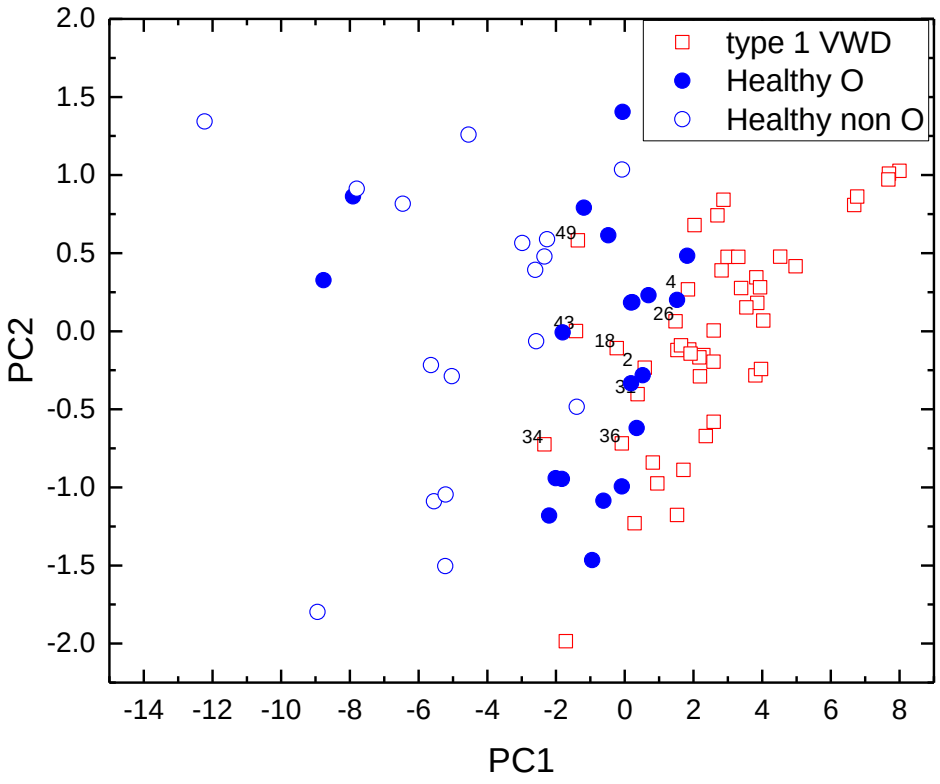


Figure 4.2 Representation of the firsts two principal component of a PCA, which capture the 93.88 % of the total variability, carried out on the experimental VWF:Ag and VWF:CB measures considering healthy and type 1 VWD subjects.

Subjects 18 and 49, even if they both carry a nonsense mutation, which usually correspond to a reduction in VWF levels, show normal VWF levels. In these subjects, in fact, the disease is extremely mild, did not affect their lifestyle, and no particular bleeding disorder were observed.

4.2 Proposed model-based algorithm for type 1 VWD diagnosis

If a subject results positive at the diagnosis algorithm presented in §1.4, specialized VWD studies are required. In Figure 4.3 is reported the proposed algorithm which, using already used specialized VWD clinical tests, aims to assist the medical work during the long procedure to obtain a type 1 VWD diagnosis.

The results shown in § 3 suggest that type 2B ad Vicenza are easily diagnosed by already developed classification method proposed by Castaldello *et al* (2017). Furthermore, the methodology used to type 2B and Vicenza classification do not interfere with algorithm proposed for type 1 VWD classification.

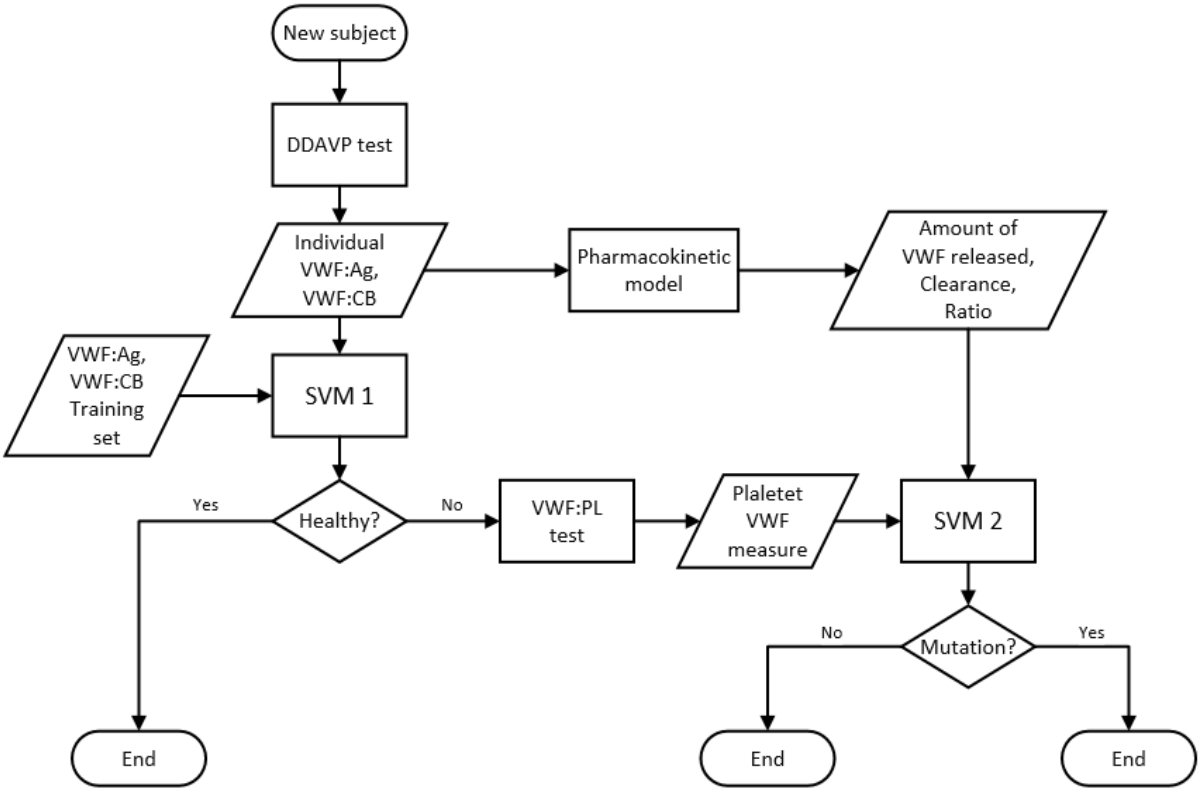


Figure 4.3 Scheme representing the proposed algorithm to obtain a type 1 VWD diagnosis depending on the presence of VWF gene mutation or not.

The proposed algorithm uses the VWF:Ag and VWF:CB levels from the DDAVP test to classify a subject as affected by VWD type 1 or healthy by support vector machine (SVM) classifier. Other methods (PLS-DA, KNN) were applied and compared but gives worst results. The second steps aims of recognising if a subject has a VWF gene mutation or not. This second classification is done using the informative pharmacokinetic indexes discussed in Chapter £ and the platelet VWF measurement. The PK indexes used are the amount of VWF released, the clearance and the ratio.

4.3 Healthy- type 1 VWD classification

Data are organized in a matrix where rows represent the subjects and columns represent the VWF:Ag and VWF:CB levels during the DDAVP test. A PCA is carried out on the organized data and the variance captured by each PC is reported in Table 4.1.

Table 4.1 Variance captured by each PC in the PCA on the DDAVP VWF levels.

PC1	PC2	PC3	PC4
90.68 %	3.20 %	1.92 %	1.57 %

A SVM classifier model with radial basis kernel function has been trained using the first three PC of the PCA, which capture a cumulative variance equal to the 95.80 % of the total. The γ and *C* SVM-classifier parameters are obtained through the leave-one-out cross validation method. The parameters that best perform in the validation activity are $\gamma = 0.2570$ and $C = 19.95$. The SVM classifier model is defined by 31 support vectors, 13 belonging to the type 1 VWD class and 18 support vectors for the healthy class.
The trained SVM model features and performances are summarized in Table 4.2 where *SV* indicate the number of support vector and *ER* the error rate in cross validation. The total error rate is calculated by the following equation (Ballabio and Todeschini, 2009):

$$ER_{tot}^{CV} = 1 - \frac{\sum_{1=1}^{G} n_{ii}}{n}; \quad (4.1)$$

where *CV* stand for cross validation activity, *G* is the number of groups and n_{ii} is the ii-th element of the confusion matrix. For the single group, the error rate is calculated considering the n_g number of subjects which belong to that group; the subscript indicate the group by which each index is related to.

Table 4.2 Summarized SVM classifier features and performances.

γ	C	SV_{tot}	$SV_{type\ 1}$	$SV_{Healthy}$	ER_{tot}^{CV}	$ER_{type\ 1}^{CV}$	$ER_{Healthy}^{CV}$
0.2570	19.95	31	13	18	0.1125	0.1190	0.1043

The total percentage misclassification error in cross validation is 11.25 %, but some subjects are wrongly classified even in the calibrated model with the optimized parameters. Subjects 36 and 49 are wrongly classified in the calibrated model; their VWF:Ag and VWF:CB levels are reported respectively in Figure 4.4a and Figure 4.4b, comparing their profiles with the average type 1 VWD and healthy blood group O subject. Both subjects shows VWF:Ag and VWF:CB levels more similar to the average healthy blood group O than the average type 1 VWD. Subject 36 carry the P2063S VWF gene mutation while subject 49 has a nonsense VWF gene mutation but he shows normal VWF levels.

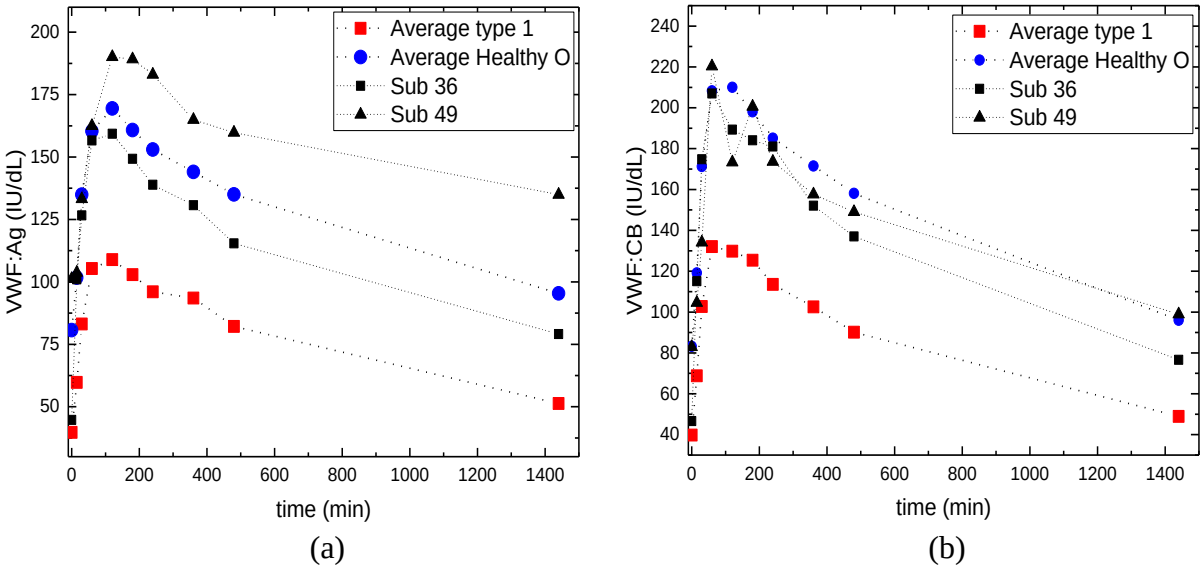


Figure 4.4 a) Experimental VWF:Ag measurement for subjects 36 and 49. b) Experimental VWF:CB measurement for subjects 36 and 49.

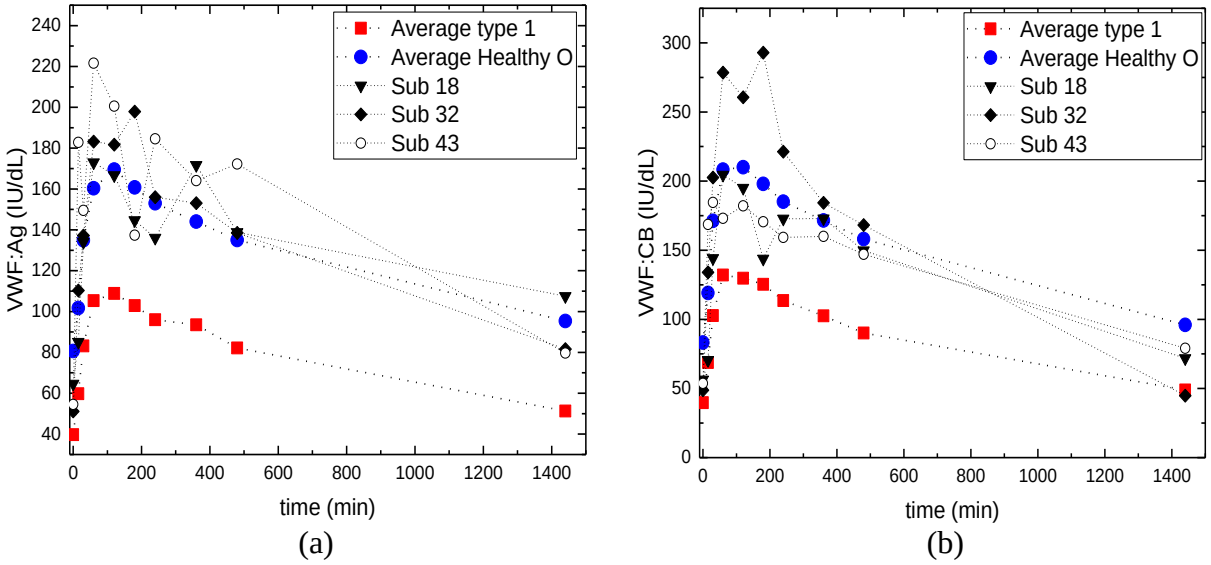


Figure 4.5 a) Experimental VWF:Ag measurement for subjects 18, 32 and 43. b) Experimental VWF:CB measurement for subjects 18, 32 and 43.

In cross-validation phase, the subjects wrongly classified are subjects 18, 32 and 43 and their VWF:Ag and VWF:CB levels are reported respectively in Figure 4.5a and Figure 4.5b. Subject 18 carry a nonsense mutation while subject 32 and 43 have no VWF gene mutation.

To make the SVM classifier more robust, a threshold probability value is applied to assign a class to a single subject. Two threshold probability values have to be chosen, one for the classification in the type 1 VWD class and one for defining the healthy class. The best couple of thresholds value are obtained through a leave-one-out method where the subject used as test can be classified as:

- type 1 VWD if its associate probability is higher than the threshold value chosen for the type 1 VWD assignment;
- healthy if its associate probability is higher than the threshold value chosen for the healthy assignment;
- type 1 VWD unclassified if the subject is classified as type 1 VWD but its associated probability is lower than the threshold value chosen for the type 1 VWD assignment;
- healthy unclassified if the subject is classified as healthy but its associated probability is lower than the threshold value chosen for the healthy assignment.

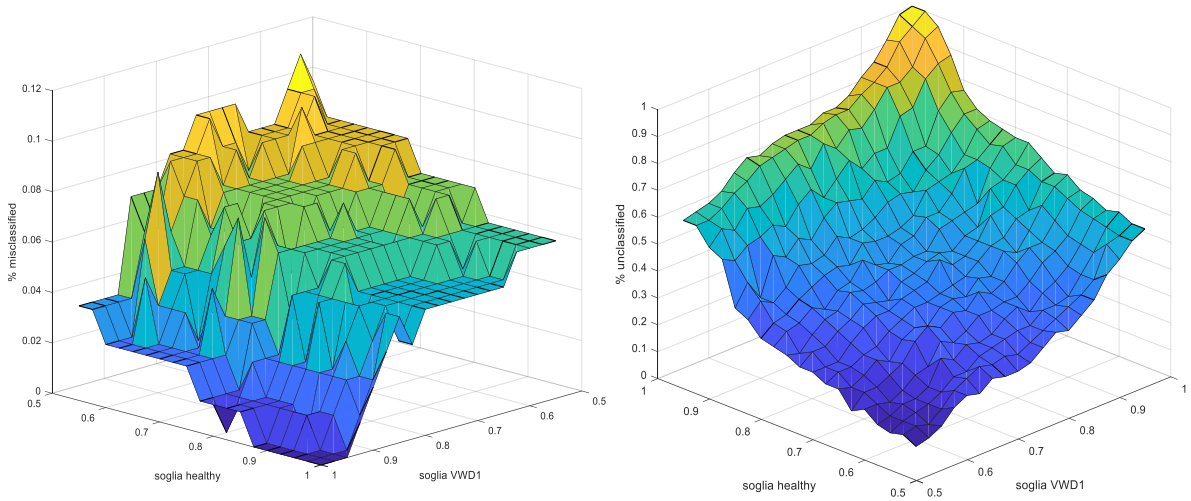
The confusion matrix obtainable with the previous classification is reported in Table 4.3.

Table 4.3 Representation of the confusion matrix obtainable form the application of thresold probability for the class assignment.

		Predicted Class		
		Healthy	**Type 1 VWD**	**Unclassified**
Real Class	**Healthy**	Healthy Classified	Healthy misclassified	Healthy unclassified
	VWD1	VWD1 misclassified	VWD1 classified	VWD1 unclassified
	Unclassified	-	-	-

.

The thresholds probability values aims to minimize the misclassified subjects. With the proposed classification criteria as the threshold values increase, the misclassification error rate decreases and the unclassified subjects increase. By looking at Figure 4.6a and Figure 4.6b, where the misclassification error rate and the unclassified percentage are reported in function of the thresholds value for each class, it can be observed that by increasing the threshold values the misclassification error decreases slower than the increase in the percentage of unclassified subjects. Consequently, the rate of subject correctly classified decreases.



(a) (b)

Figure 4.6 a) Representation of the error rate of the SVM classifier depending on the threshold probability value for each class assignment. b) Representation of the percentage of unclassified subjects depending on the threshold probability value for each class assignment.

The error rate and the unclassified subjects percentage for the thresholds values comprised between 0.550 and 0.650 are reported respectively in Table 4.4 and Table 4.5 for the type 1 VWD class and in Table 4.6 and Table 4.7 for the healthy class.

Table 4.4 Misclassified error rate for type 1 VWD subjects depending on threshold probability values comprises between 0.550 and 0.650. Bold value corresponds to the chosen pair of threshold probability value.

		Healthy probability threshold				
		0.550	**0.575**	**0.600**	**0.625**	**0.650**
Type 1 VWD probability threshold	**0.550**	0.1190	0.1190	0.1190	0.1190	0.0952
	0.575	0.1190	0.1190	0.1190	0.1190	0.1190
	0.600	0.1190	0.1190	0.1190	0.1190	**0.0952**
	0.625	0.1190	0.1190	0.1190	0.0952	0.1190
	0.650	0.1190	0.1190	0.0952	0.0952	0.0952

Table 4.5 Unclassified rate for type 1 VWD subjects depending on threshold probability values comprises between 0.550 and 0.650. Bold value corresponds to the chosen pair of threshold probability value.

		Healthy probability threshold				
		0.550	**0.575**	**0.600**	**0.625**	**0.650**
Type 1 VWD probability threshold	**0.550**	0.0238	0.0238	0.0238	0	0.0476
	0.575	0.0238	0.0238	0.0476	0.0238	0.0238
	0.600	0.0238	0.0238	0.0238	0.0238	**0.0476**
	0.625	0.0714	0.0714	0.0714	0.0476	0.0476
	0.650	0.0714	0.0714	0.0952	0.1190	0.0714

Table 4.6 Misclassified error rate for healthy subjects depending on threshold probability values comprises between 0.550 and 0.650. Bold value corresponds to the chosen pair of threshold probability value.

		Healthy probability threshold				
		0.550	**0.575**	**0.600**	**0.625**	**0.650**
Type 1 VWD probability threshold	**0.550**	0.0750	0.1000	0.0750	0.0750	0.1000
	0.575	0.0750	0.0750	0.0750	0.0750	0.0750
	0.600	0.0500	0.0750	0.0500	0.0750	**0.0750**
	0.625	0.0500	0.0500	0.0500	0.0750	0.0750
	0.650	0.0500	0.0500	0.0500	0.0500	0.0500

Table 4.7 Unclassified rate for healthy subjects depending on threshold probability values comprises between 0.550 and 0.650. Bold value corresponds to the chosen pair of threshold probability value.

		Healthy probability threshold				
		0.550	**0.575**	**0.600**	**0.625**	**0.650**
Type 1 VWD probability threshold	**0.550**	0.1000	0.1000	0.2000	0.2000	0.2250
	0.575	0.0750	0.1500	0.1500	0.1500	0.2250
	0.600	0.1250	0.1250	0.1500	0.2000	**0.2250**
	0.625	0.1250	0.1250	0.2000	0.2250	0.2250
	0.650	0.1500	0.1500	0.2000	0.2250	0.2500

Table 4.8 Total misclassified rate depending on threshold probability values comprises between 0.550 and 0.650. Bold value corresponds to the chosen pair of threshold probability value.

		Healthy probability threshold				
		0.550	**0.575**	**0.600**	**0.625**	**0.650**
Type 1 VWD probability threshold	**0.550**	0.1000	0.1125	0.1000	0.1000	0.1000
	0.575	0.1000	0.1000	0.1000	0.1000	0.1000
	0.600	0.0875	0.1000	0.0875	0.1000	**0.0875**
	0.625	0.0875	0.0875	0.0875	0.0875	0.1000
	0.650	0.0875	0.0875	0.0750	0.0750	0.0750

Table 4.9 Total unclassified rate depending on threshold probability values comprises between 0.550 and 0.650. Bold value corresponds to the chosen pair of threshold probability value.

		Healthy probability threshold				
		0.550	**0.575**	**0.600**	**0.625**	**0.650**
Type 1 VWD probability threshold	**0.550**	0.6000	0.6000	0.0719	0.0950	0.1319
	0.575	0.0481	0.0838	0.1081	0.0838	0.1194
	0.600	0.0719	0.0719	0.0838	0.1075	**0.1319**
	0.625	0.0969	0.0969	0.1325	0.1319	0.1319
	0.650	0.1087	0.1087	0.1450	0.1694	0.1563

Table 4.8 and Table 4.9 report the total misclassification error rate and the total percentage of subjects unclassified for the threshold values comprised between 0.550 and 0.650. The lower total misclassification error rate corresponds to the higher unclassified percentage; the threshold probability values are chosen searching for the lower misclassification error with a reasonable unclassified percentage. With a probability threshold of 0.650 for the healthy class assignment and 0.600 for the type 1 VWD assignment, the misclassification error rate is 0.0875 and the unclassified rate is 0.1319. The SVM classifier shows an error rate for type 1 VWD class assignment equal to 0.0952 (Table 4.4), which is slightly higher than the one for the healthy class equal to 0.0750 (Table 4.6). The unclassified rates are quite different for the two classes; for the type 1 VWD class, the unclassified rate is 0.0476 (Table 4.5), while for the healthy class is 0.225 (Table 4.7). These results have an important meaning; if a subject is VWD type 1 affected, the classifier assign the VWD type 1 class with a high confidence, while if a subject is healthy, the classifier shows a relative high uncertainty to assign the healthy class.
This reflects the difficulties related to the type 1 VWD recognition from the healthy subjects.

4.4 VWD type 1 data analysis for classification

When a subject is classify as affected by type 1 VWD, the second step aims at the classification depending on the presence of a VWF gene mutation or not.

At this scope, a SVM classifier is trained using a training dataset composed by 41 subjects; 30 are the subjects with VWF gene mutations comprising missense and nonsense mutation and 11 are the subjects without mutations.

The most informative indexes are the amount of VWF released and the clearance obtained from the PK model, the ratio between the VWF:Ag and VWF:CB measures at the basal state and the measure of platelet VWF. The platelet VWF measurement is introduced to take into account the VWF synthesis independently of other factors that instead affect the plasma VWF levels (Casonato *et al*, 2016). Since the VWF synthesis mechanism is consider equally in bone narrow and in subendothelium, low platelet VWF levels is associated to problems related to its synthesis, which can be difficultly detected measuring only the VWF released during the DDAVP test.

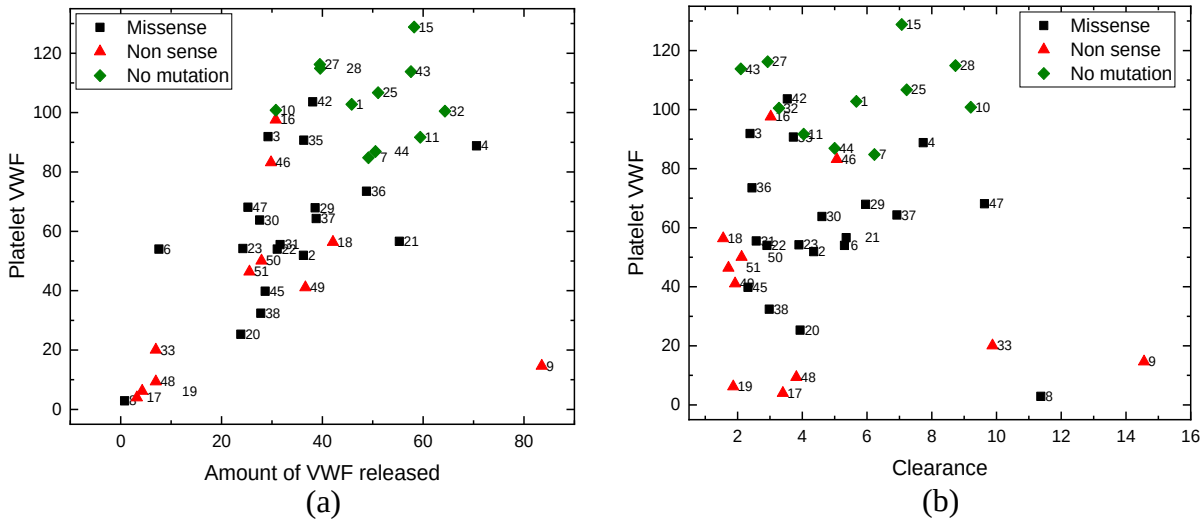


Figure 4.7 a) Amount of VWF released and platelet VWF for each subjects affected by type 1 VWD. b) Clearance and platelet VWF for each subjects affected by type 1 VWD.

In Figure 4.7a, the platelet VWF level and the amount of VWF released are reported for each subject, while in Figure 4.7b the platelet VWF level and clearance are reported for each subject. Subjects without gene mutation show an average platelet VWF level higher than the subjects with gene mutation with some exception for both missense and nonsense mutation subjects. Furthermore, the severe type 1 VWD subjects show a platelet VWF value, which is lower than the mild type 1 VWD subjects.

4.5 Mutation-no mutation classification

The data discussed in the previous paragraph are not directly used are predictors to train the SVM classifier, but a PCA is carried out on the amount of VWF released, clearance, ratio and platelet data to reduce the dimensionality and avoid noise measurement; the variance captured by each PC is reported in Table 4.10.

Table 4.10 Captured percentage variance for each PC for the PCA carry out on the platelet VWF, the amount of VWF released, the clearance and the ratio data for the subjects affected by the type 1 VWD.

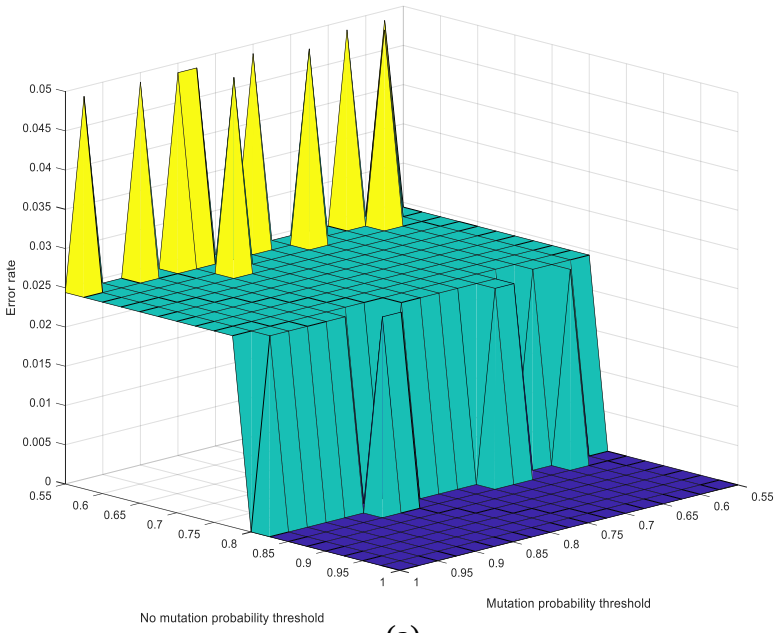
PC1	PC2	PC3
45.38 %	25.77 %	19.07 %

A SVM classifier with radial basis function is trained on the firsts three PC which capture a cumulative variance equal to the 90.2 % of the total. The SVM parameters are obtained through a leave-one-out cross-validation and are reported in Table 4.11. In Table 4.11 also the number of support vector needed for each class are reported.

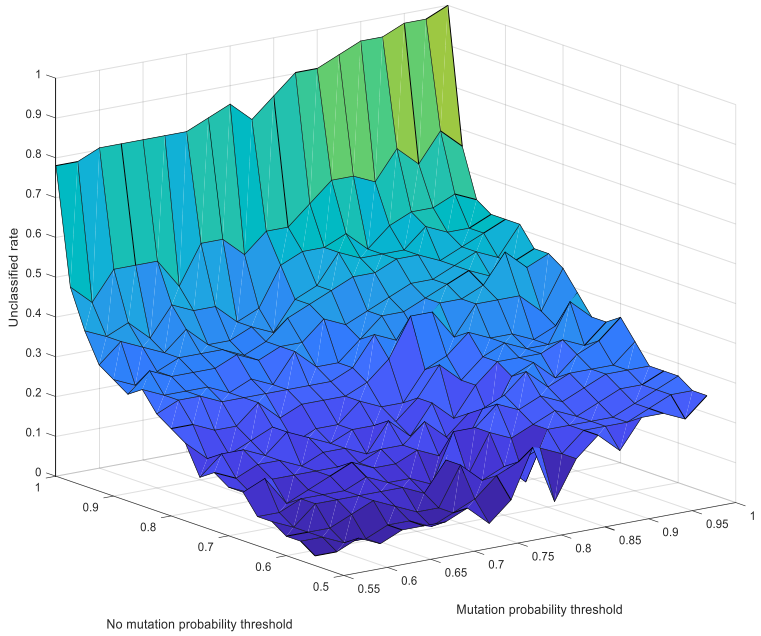
Table 4.11 Summarized SVM classifier features and performances.

γ	C	SV_{tot}	SV_{Mut}	$SV_{No\ Mut}$	ER_{tot}^{CV}	ER_{Mut}^{CV}	$ER_{No\ Mut}^{CV}$
0.3548	7.08	15	10	5	0.0244	0	0.091

In cross-validation, the misclassification total error rate is 0.0244 (Table 4.11). Looking at each class, no subjects with VWF gene mutation are wrongly classified, while a single subject without VWF gene mutation is wrongly classified.
With the same procedure adopted for the classification between healthy and type 1 VWD class, to increase the classifier robustness, a threshold probability is applied for each class assignment. In Figure 4.8a the misclassification error rate and the unclassified rate in function of the probability thresholds values for each class is shown. In Figure 4.8a the independency of the thresholds value with respect to the mutation class misclassification error is evident; in fact no misclassified subjects are present in cross validation. In Figure 4.8b the unclassified rate is reported in function of the probability thresholds values for each class and increases for both classes when the threshold value, for each class, increases.



(a)



(b)

Figure 4.8 a) Representation of the total error rate of the SVM classifier depending on the threshold probability value for each class assignment (mutation or no mutation). b) Unclassified rate representation depending on the threshold probability value for each class assignment.

Table 4.12 Misclassification rate for the no mutation class depending on threshold probability values comprises between 0.550 and 0.650. Bold value corresponds to the chosen pair of threshold probability value.

		Mutation probability threshold					
		0.550	**0.570**	**0.590**	**0.610**	**0.630**	**0.650**
No mutation probability threshold	**0.550**	0.0909	0.0909	0.1818	0.0909	0.0909	0.1818
	0.570	0.0909	0.0909	0.0909	0.0909	0.0909	0.1818
	0.590	0.0909	0.0909	0.0909	0.0909	0.1818	0.0909
	0.610	0.0909	0.0909	0.0909	0.0909	0.0909	0.0909
	0.630	**0.0909**	0.0909	0.0909	0.0909	0.0909	0.0909
	0.650	0.0909	0.0909	0.0909	0.0909	0.0909	0.0909

Table 4.13 Unclassified rate for the no mutation class depending on threshold probability values comprises between 0.550 and 0.650. Bold value corresponds to the chosen pair of threshold probability value.

		Mutation probability threshold					
		0.550	**0.570**	**0.590**	**0.610**	**0.630**	**0.650**
No mutation probability threshold	**0.550**	0.0909	0.1818	0.0909	0.1818	0.1818	0.0909
	0.570	0.0909	0.1818	0.1818	0.1818	0.1818	0.0909
	0.590	0.1818	0.1818	0.1818	0.1818	0.0909	0.1818
	0.610	0.0909	0.1818	0.1818	0.1818	0.1818	0.1818
	0.630	**0.0909**	0.1818	0.1818	0.1818	0.1818	0.1818
	0.650	0.1818	0.1818	0.1818	0.1818	0.0909	0.1818

Table 4.14 Unclassified rate for the mutation class depending on threshold probability values comprises between 0.550 and 0.650. Bold value corresponds to the chosen pair of threshold probability value.

		Mutation probability threshold					
		0.550	**0.570**	**0.590**	**0.610**	**0.630**	**0.650**
No mutation probability threshold	**0.550**	0	0	0	0	0	0
	0.570	0	0.0333	0	0	0	0.0333
	0.590	0	0	0.0333	0	0	0.0333
	0.610	0	0.0333	0.0333	0	0.0333	0.0333
	0.630	**0.0333**	0.0667	0.0333	0.0667	0.0333	0.0333
	0.650	0.0333	0.0333	0.0667	0.0333	0	0.0667

Table 4.15 Total unclassified rate depending on threshold probability values comprises between 0.550 and 0.650. Bold value corresponds to the chosen pair of threshold probability value.

		Mutation probability threshold					
		0.550	**0.570**	**0.590**	**0.610**	**0.630**	**0.650**
No mutation probability threshold	**0.550**	0.0244	0.0488	0.0244	0.0488	0.0488	0.0244
	0.570	0.0244	0.0732	0.0488	0.0488	0.0488	0.0488
	0.590	0.0488	0.0488	0.0732	0.0488	0.0244	0.0732
	0.610	0.0244	0.0732	0.0732	0.0488	0.0732	0.0732
	0.630	**0.0488**	0.0976	0.0732	0.0976	0.0732	0.0732
	0.650	0.0732	0.0732	0.0976	0.0732	0.0244	0.0976

In Table 4.15 the total unclassified rate, which comprise both classes for the threshold values between 0.550 and 0.650, is reported. The misclassification error rate in this range is constant at the values of 0.0244, meaning that a single subject is wrongly classified. With a threshold probability of 0.550 for the mutation category and 0.630 for the no mutation category, the total unclassified rate is 0.0488, which means that two subjects are not assigned to a class. Looking at the single class unclassified rate (Table 4.13 and Table 4.14), only a subject without VWF gene mutation and one with VWF gene mutation are not assigned to the belonging class. The misclassification rate is zero for the mutation class and is 0.09 for the no mutation class (Table 4.12). Subject 27 is the one wrongly classified. It shows a high platelet VWF value but levels of amount of VWF released and clearance more similar to the missense and nonsense VWF gene mutation subjects. The unclassified subject for the mutation class is subject 42, which carry a missense mutation and shows a higher platelet VWF than the other missense subjects. The unclassified subject without gene mutation is the number 44, which has a platelet VWF level at the boundary between the two categories.

Conclusions

Von Willebrand disease diagnosis is a complex task requiring many clinical test to obtain a diagnosis. Because of the many VWD variants and the high heterogeneity observed within the same variant the DNA sequencing is often needed to confirm the diagnosis. Type 1 VWD is the most common VWD type, covering about the 65% of the total cases and its behaviour is very similar to subjects belonging to healthy O blood type.

A pharmacokinetic model, which was developed to describe the release, proteolysis and the clearance of VWF multimers following the injection of DDAVP, has been used to characterized the behaviours of subjects affected by type 1 VWD. A clinical dataset if 51 subjects was used for the model identification.

The pharmacokinetic model well represents the general behaviour of the analysed subjects with respect the other VWD types. The pharmacokinetic model well characterizes the intra type 1 variants, showing a clear distinction between the subjects with severe and mild type 1 VWD. A bivariate analysis carried out on the amount of VWF released and the clearance shows the presence of two clusters; one that identify the subjects with VWF gene mutation (namely missense and nonsense) and one that identify mainly subjects without VWF gene mutation.

In order to achieve a good classification among type 1 and healthy subjects, a SVM classifier has been trained using the raw VWF:Ag and VWF:CB data. The SVM classifier has been validated with a cross-validation method and 80% of the subjects are correctly classified with a high confidence. The remaining 20% of the subjects are equally divided in subjects correctly classified but with low confidence and subjects wrongly classified. The classifier shows a higher uncertainty for the healthy class assignment respect to the type 1 class, suggesting that the affected subjects are more easily recognizable.

Once a subject is classified as affected by type 1 VWD, a new classification task must be performed in order to assess the correct type 1 variant. With this aim, a new SVM classifier has been trained using some indexes obtained from the pharmacokinetic model, namely the amount of VWF released and the clearance, together with the platelet VWF experimental measurements and the VWF ratio. In particular, the classifier identifies if a subject carry or not a gene mutation linked to the disease. The SVM classifier has been validated through a cross-validation method and correctly classifies with high confidence more than 90% of the subjects. Only 2.5% of the subjects are wrongly classified, while 7.5% are correctly classified but with low confidence.

With the aforementioned two-step classification based on the available historical database, it is possible to build an algorithm that is suitable to assist expert clinical practitioners in the classification of a new subject.

To improve the proposed algorithm a new classification step should be included to distinguish between subjects with nonsense mutation and missense mutation, but taking into account other variables since from DDAVP data and PK indexes this does not appear to be possible.

References

Ballabio, D., Todeschini, R., (2009). Multivariate classification for qualitative analysis. *Infrared spectroscopy for food quality analysis and control*, Elsevier, Burlington, U.S.A., p. 83-104.

Bard, Y. (1974). *Nonlinear parameter estimation*. Academy Press, New York, U.S.A.

Berntorop, E. (2007). Erik von Willebrand. *Thrombosis research,* **120**: S3-S4.

Casonato, A., Cattini, M.G., Daidone, V., Pontara, E., Bertomoro, A., Prandoni, P., (2016). Diagnostic value of measuring platelet von Willebrand factor in von Willebrand disease. *PLoS ONE* **11**:8.

Castaldello, C., Galvanin, F., Casonato, A., Padrini, R., Barolo, M., Bezzo, F., (2017). A model-based protocol for the diagnosis of von Willebrand disease. *Canadian Journal of Chemical Engineering*, **xx**:1-10.

Emery, A. F., (2001). Using the concept of information to optimally design experiments with uncertain parameters. *ASME Journal of Heat Transfer*, **123**: 593-600.

Esbensen, H., Geladi, P. (2009). Principal Component Analysis: Concept, Geometrical Interpretation, Mathematical Background, Algorithms, History, Practice. Comprehensive Chemometrics.

Ferrari, M., Galvanin, F., Barolo, M., Daidone, V., Padrini, R., Bezzo, F., Casonato, A. (2018). A mechanistic model to quantify von Willebrand factor release, survival and proteolysis in patients with von Willebrand disease. *Thrombosis and Haemostasis*, **118**: 309-319.

Galvanin, F. (2010*). Optimal model-based design of experiments in dynamic system: novel techniques and unconventional applications.* PhD Thesis, Università degli Studi di Padova, Italy.

Galvanin, F., Barolo, M., Padrini, R., Casonato, A., Bezzo, F. (2014). A model-based approach to the automatic diagnosis of von Willebrand disease. *American Institute of Chemical Engineers*, **60**:1718-1727.

Hampshire, D., J., Goodeve, A., C., (2013). P.P2063S: a neutral VWF variant masquerading as a mutation. *Haemostasis research,* **93**:505-506.

James, G., Witten, D., Hastie, H., Tibshirani, R., (2013). *An Introduction to Statistical Learning*. Springer, New York, p. 337-353.

Kumar, V., Abbas, A.K., Aster, J. C. (2015). Robbins and Cotran – *Pathologic basis of disease.* Elsevier, Philadelphia, U.S.A., 9^{th} edition.

Lillicrap, D. (2007). Von Willebrand disease-phenotype versus genotype: Deficiency versus didease. *Thrombosis Research,* **87**: 57-64.

Menache, D., Aronson, D., Darr, F., Montgomery, R., Gill, J., Kessler, C., Lusher, J., phatak, P., Shapiro, A., Thompson, A., White, G. (1996). Pharmacokinetics of von Willebrand factor and factor VIIIc in patients with severe von Willebrand disease (type 3 VWD): Estimaion of the rate of factor VIIc synthesis. *British Journal of Haematology*, **94**: 740-745.

N.H.B.L.I. (2007). *The Diagnosis, Evaluation and Management of von Willebrand Disease.* Technical Report 08-5832, US Department of Health and Human Services – National Institute of Health – National Heart, Lung and Blood Institution, Bethesda, U.S.A.

Platt, C., (1999). Probabilistic outputs for support vector machines and comparison to regularized likelihood methods. *Advances in Large margin Classifiers*, *MIT Press*.

Process System Enterprise, (2012a). gPROMS Model Developer Guide, v. 3.6. Process system Enterprise Ltd, London, UK.

Process System Enterprise, (2012b). gPROMS Model Validation Guide, v. 3.6. Process system Enterprise Ltd, London, UK.

Sadler, J.E., Manucci, P.M., Berntorp, E., Bochkov, N., Boulyjenkov, V., Ginsburg, V., Meyer, D., Peake, D., Rodeghiero, F., Srivastava, A. (2000) Impact, diagnosis and treatment of von WIllebrand disease. *Thrombosis and haemostasis*, **84**:160-74.

Sadler, J. E. (2003). Von Willebrand disease type 1: a diagnosis in search of a disease. *Blood Journal,* **101**: 2089-2093.

Sadler, J. E. (2005). Von Willebrand factor: two side of a coin. *Thrombosis and Haemostasis*, **3**:1702-1709.

Taverna, B. (2017). *Optimal design of a minimum set of clinical tests for the identification and characterization of von Willebrand disease.* Tesi di Laurea Magistrale, Dipartimento di Ingegneris Industriale, Università degli Studi di Padova.

Wise, B., Gallagher, N.B., (1996). The process chemometrics approach to process monitoring and fault detection. *Journal of Process Control*, **6**: 329-348.

Xu, Y. Zomer, S., Brereton, R.G. (2006). Support vector machines: a recent method for classification in chemometrics. *Critical Reviews in Analytical Chemistry*, **36**(3-4): 177-188.

Zaverio, R.M. (1999). Structure and function of von Willebrand Factor. *Thrombosis and Haemostasis*, **82**: 576-584.

Zaverio, M. (2007). The role of von Willebrand factor in thrombus formation. *Thrombosis Research*, **120**: S9-S16.

Zullo, L. C. (1991). *Computer aided design of experiments. An engineer approach*. PhD Thesis, University of London, UK.