Journal of Pharmaceutical and Biomedical Analysis 54 (2011) 1020–1029

Contents lists available at ScienceDirect

Journal of Pharmaceutical and Biomedical Analysis

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

Combining ^{1}H NMR spectroscopy and chemometrics to identify heparin samples that may possess dermatan sulfate (DS) impurities or oversulfated chondroitin sulfate (OSCS) contaminants

Qingda Zang[a,b,c], David A. Keire[d], Richard D. Wood[b], Lucinda F. Buhse[d], Christine M.V. Moore[e], Moheb Nasr[e], Ali Al-Hakim[e], Michael L. Trehy[d], William J. Welsh[a,*]

[a] Department of Pharmacology, Robert Wood Johnson Medical School, University of Medicine & Dentistry of New Jersey, Piscataway, NJ 08854, USA
[b] Snowdon, Inc., 1 Deer Park Drive, Suite H-3, Monmouth Junction, NJ 08852, USA
[c] Department of Health Informatics, School of Health Related Professions, University of Medicine & Dentistry of New Jersey, Newark, NJ 07107, USA
[d] Food and Drug Administration, CDER, Division of Pharmaceutical Analysis, St Louis, MO 63101, USA
[e] Food and Drug Administration, CDER, Office of New Drug Quality Assessment, Silver Spring, MD 20993, USA

ARTICLE INFO

Article history:
Received 5 September 2010
Received in revised form 8 November 2010
Accepted 10 December 2010
Available online 16 December 2010

Keywords:
Heparin
Proton nuclear magnetic resonance spectroscopy
Pattern recognition
Principal components analysis
Linear discriminant analysis

ABSTRACT

Heparin is a naturally produced, heterogeneous compound consisting of variably sulfated and acetylated repeating disaccharide units. The structural complexity of heparin complicates efforts to assess the purity of the compound, especially when differentiating between similar glycosaminoglycans. Recently, heparin sodium contaminated with oversulfated chondroitin sulfate A (OSCS) has been associated with a rapid and acute onset of an anaphylactic reaction. In addition, naturally occurring dermatan sulfate (DS) was found to be present in these and other heparin samples as an impurity due to incomplete purification. The present study was undertaken to determine whether chemometric analysis of these NMR spectral data would be useful for discrimination between USP-grade samples of heparin sodium API and those deemed unacceptable based on their levels of DS, OSCS, or both. Several multivariate chemometric methods for clustering and classification were evaluated; specifically, principal components analysis (PCA), partial least squares discriminant analysis (PLS-DA), linear discriminant analysis (LDA), and the *k*-nearest-neighbor (kNN) method. Data dimension reduction and variable selection techniques, implemented to avoid over-fitting the training set data, markedly improved the performance of the classification models. Under optimal conditions, a perfect classification (100% success rate) was attained on external test sets for the Heparin vs OSCS model. The predictive rates for the *Heparin* vs *DS*, *Heparin* vs [*DS* + *OSCS*], and *Heparin* vs *DS* vs *OSCS* models were 89%, 93%, and 90%, respectively. In most cases, misclassifications can be ascribed to the similarity in NMR chemical shifts of heparin and DS. Among the chemometric methods evaluated in this study, we found that the LDA models were superior to the PLS-DA and kNN models for classification. Taken together, the present results demonstrate the utility of chemometric methods when applied in combination with ^{1}H NMR spectral analysis for evaluating the quality of heparin APIs.

1. Introduction

Starting in November 2007, hundreds of cases of adverse reactions to heparin, such as hypotension, severe allergic symptoms, and even death, were reported to the US Food and Drug Administration (FDA) [1]. Prompted by these adverse events, biological and analytical methods were developed to identify contaminants and impurities in heparin [2–6]. Oversulfated chondroitin sulfate (OSCS) was identified as a contaminant associated with these adverse clinical effects [7]. Heparin is a polydisperse linear polysaccharide consisting primarily of alternating glucosamine and hexuronic acid with various sulfonated and acetylated substitutions [8]. Like heparin, OSCS is a polyanionic glycosaminoglycan but differs in structure. In OSCS, the 1,3 linked disaccharide units are sulfated at the 4-*O* and 6-*O* positions of the galactosamine as well as at the 2-*O* and 3-*O* positions of the glucuronic acid. OSCS is not known to be a natural product; it can be synthesized by chemically modifying chondroitin sulfate A (CSA), which normally contains one sulfate group per disaccharide unit [9]. In standard drug potency assays, the OSCS molecule can partially mimic the anti-coagulation activity of heparin [10]. In addition, a naturally

* Corresponding author at: Department of Pharmacology, Robert Wood Johnson Medical School, University of Medicine & Dentistry of New Jersey, 661 Hoes Lane West, Room SRB-125, Piscataway, NJ 08854, USA. Tel.: +1 732 235 3234; fax: +1 732 235 3475.
E-mail address: welshwj@umdnj.edu (W.J. Welsh).

0731-7085/$ – see front matter © 2010 Elsevier B.V. All rights reserved.
doi:10.1016/j.jpba.2010.12.008

occurring polysaccharide, dermatan sulfate (DS, formerly named chondroitin sulfate B or CSB), can be present in heparin products due to incomplete purification [11,12].

To ensure the safety and quality of heparin, spectroscopy and chromatography methods have been added to the United States Pharmacopeia (USP) monograph for heparin sodium active pharmaceutical ingredient (API) to detect and screen for impurities and contaminants [13]. Nuclear magnetic resonance (NMR) spectroscopy is now used to identify the presence or absence of contaminants [7,9,11,14], while capillary electrophoresis (CE) [6,12,15] and strong anion exchange-HPLC (SAX-HPLC) [5,16,17] have been used to measure the relative amounts of heparin, DS and OSCS. SAX-HPLC and CE assays on heparin samples from 2008 found up to 27% (w/w%) OSCS and 19% DS. However, of these three analytical techniques, the complex pattern of overlapping ^{1}H NMR signals found in the heparin spectra was judged most effective to assess structural information.

Because unique signals associated with OSCS or DS in contaminated or impure heparin were observed in the NMR spectra, the present study was undertaken to evaluate the ability of several chemometric approaches, including principal components analysis (PCA), partial least squares discriminant analysis (PLS-DA), linear discriminant analysis (LDA), and the *k*-nearest neighbor (kNN) method to distinguish between pure, impure or contaminated samples of heparin based on analysis of their ^{1}H NMR spectral data. To maintain consistency in data handling and to avoid bias, blinded ^{1}H NMR data from heparin samples analyzed by FDA personnel was provided for subsequent chemometric analysis. The purpose of the study was to assess the ability of these chemometric approaches to differentiate the samples into distinct groups corresponding to pure, impure, or contaminated heparin based solely on analysis of their NMR spectral data and, more generally, to characterize analytes for quality control or purity assessment. Results from the present study demonstrate success rates of 90–100% classification using chemometric approaches, with LDA and PLS-DA providing the best performance overall.

2. Materials and methods

2.1. NMR spectroscopy measurement

All samples were analyzed using a Varian Inova 500 instrument at the Washington University (St. Louis, MO) Chemistry Department NMR Facility operating at 499.893 MHz for ^{1}H-nuclei. Details on the NMR measurements of the heparin samples can be found elsewhere [18].

2.2. HPLC analysis of heparin

SAX-HPLC was used to measure the weight percent of DS or OSCS content in heparin as described in previous work [5,13]. Briefly, an Agilent 1100 HPLC system with a Dionex AS-11HC strong anion exchange column was used to separate DS, heparin and OSCS. The peak areas of DS (eluting at 16 min) or OSCS (eluting at 24 min) were compared to the areas obtained from the response of DS or OSCS standards to quantify the weight percent of the impurity and contaminant in the presence of heparin. In addition, the heparin samples were analyzed using HPLC with a pulsed amperometric detector as described in the USP monograph [19] to determine the weight percent galactosamine (%Gal) content. This assay measures total galactosamine content and does not discriminate between galactosamine containing impurities or contaminants (*e.g.* DS, CSA or OSCS).

2.3. Data processing

^{1}H NMR spectra for over 170 heparin sodium API samples from different manufacturers with varying levels of OSCS and DS were processed using the software MestRe-C (Version 5.3.0). Phase correction was achieved through automatic zero and first-order correction procedures. All samples were made *ca.* 3 mM in 4,4-dimethyl-4-silapentane-1-sulfonic (DSS) acid as an internal reference for chemical shift.

For the chemometric analysis, each ^{1}H NMR spectrum was automatically data-reduced and converted into 125 variables by dividing the 1.95–5.70 ppm region into sequential windows of width 0.03 ppm. After exclusion of the windows containing signals due to residual processing solvents and reagents, a reduced data set of 74 variables was obtained for subsequent data analysis. Prior to chemometric analysis, the spectra were converted into ASCII files where the data were represented in $n \times m$-dimensional space (n and m equal to the number of samples and the number of variables, respectively), and the resulting data matrix was imported into Microsoft Excel 2003. Following standard practice in multivariate chemometric analysis, the total data set was divided into two subsets: a training set for building and calibrating the classification models; and a test set for validating the model's predictive ability.

Although the current USP monograph specifies the weight percent of galactosamine (%Gal) may not exceed 1% in total hexosamine content, the pending Stage 3 revision of the heparin sodium monograph requested by the FDA specifies 1.0%Gal. Therefore, the 1.0%Gal specification was adopted in the present study to delineate heparin samples that do or do not pass this criterion. DS is the primary chondroitin impurity observed in heparin APIs and, for the purpose of this study, the %Gal is presumed equal to %DS for samples not containing OSCS. The samples were divided into three groups: (a) *Heparin*: DS $\leq$ 1.0% and OSCS = 0%; (b) *DS*: DS > 1.0% and OSCS = 0%; and (c) *OSCS*: OSCS > 0% with any content of DS. The total data set comprised 178 samples, consisting of 82 *Heparin*, 50 *DS*, and 46 *OSCS* samples.

2.4. Software

All data processing, multivariate analysis, and model building were implemented using the *R* statistical analysis software for Windows (Version 2.8.1) [20]. The packages *stats*, *caret*, *MASS*, as well as *class* and *chemometrics* in *R* were used to perform principal components analysis, partial least squares discriminant analysis, linear discriminant analysis, and *k*-nearest neighbors analysis, respectively [21,22]. The variable selection through stepwise linear discriminant analysis (SLDA) was conducted using the chemometric software *V-Parvus 2007* [23].

3. Results and discussion

3.1. Proton NMR spectra

Representative ^{1}H NMR spectra of the three classes of heparin samples (*i.e.*, *Heparin*, *DS* and *OSCS*) are illustrated in Fig. 1 for the range 1.95–6.0 ppm. Each spectrum reveals distinctive features, and their respective patterns are easily distinguished from one other in the range from 1.95 to 2.20 ppm. The basic repeating disaccharide unit for heparin is 2-O-sulfated uronic acid and 6-O-sulfated N-sulfated glucosamine, whereas the corresponding repeating unit for DS or OSCS is iduronic or glucuronic acid, respectively, and galactosamine. About every fifth amino group is acetylated for heparin, but almost all of the amino groups are acetylated in DS and OSCS [7,11]. A single peak appears at 2.05 ppm for the N-acetyl protons of heparin, and the methyl sig-

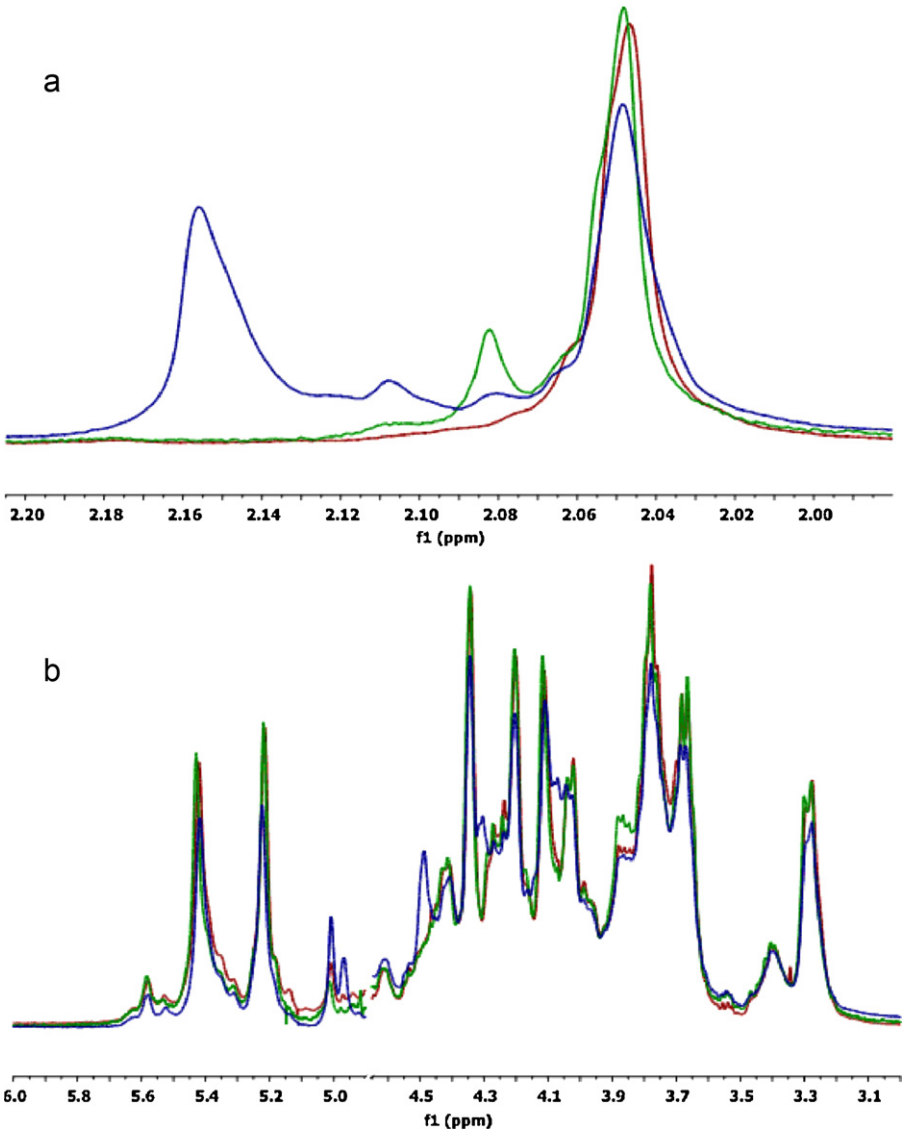

Fig. 1. ^{1}H NMR spectra of pure heparin (brown, 0.14% DS and 0% OSCS), heparin containing DS impurity (green, 4.12% DS and 0% OSCS), and heparin containing OSCS contaminant (blue, 2.71% DS and 14.0% OSCS). (a) In the 2.20–1.95 ppm region; (b) in the 6.00–3.00 ppm region. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of the article.)

nal shifts about 0.03 ppm downfield in *DS* samples. Thus, a small peak, corresponding to the N-acetyl protons of DS, can be found near 2.08 ppm. Likewise, OSCS exhibits a characteristic peak near 2.15 ppm. Other less obvious differences in these three classes of heparin samples occur in the remaining pattern of intensities at the chemical shifts of the heparin protons in the NMR spectra. These patterns of intensities are valuable for characterizing and quantifying analytes for quality control and purity assessment [24–26], and amenable to analysis using chemometric approaches. Prior to chemometric analysis, the ^{1}H NMR spectra of the heparin samples were preprocessed into a discrete set of variables that served as the input to the pattern recognition tools for subsequent analysis of the pure, DS-impure, and OSCS-contaminated heparin samples.

3.2. Pattern recognition

For the classification studies, the 178 heparin samples were divided into three groups, viz., *Heparin:* pure heparin with DS $\leq$ 1.0% and OSCS = 0%; *DS*: impure heparin with DS > 1.0% and OSCS = 0%; and *OSCS*: contaminated heparin with OSCS > 0% and with any content of DS. In these samples, the DS content varied from 0% to19% of the disaccharide mixture, and the OSCS content varied from 0% to 27%. The data set of 178 heparin samples was split (2:1) into 118 samples for training (54 *Heparin*, 33 *DS*, and 31 *OSCS*) and 60 samples for external validation and testing (28 *Heparin*, 17 *DS*, and 15 *OSCS*). The ^{1}H NMR spectral data for the data set were represented as a two-dimensional array with each row corresponding to a sample and the columns to the 74 variables.

3.2.1. Principal components analysis (PCA)

PCA is a well-known technique for reducing the dimensionality and simplifying the visualization of complex multivariate data sets [26,27]. PCA transforms the original variables into a smaller number of mutually orthogonal variables called principal components (PCs). The first component (PC1) explains the maximum amount of variance in the data, followed by PC2, next PC3, and so on. PCA is an unsupervised method, in that no *a priori* knowledge relating to class affiliation is required [28]. PCA has been widely applied in conjunction with discriminant analysis techniques to handle clas-

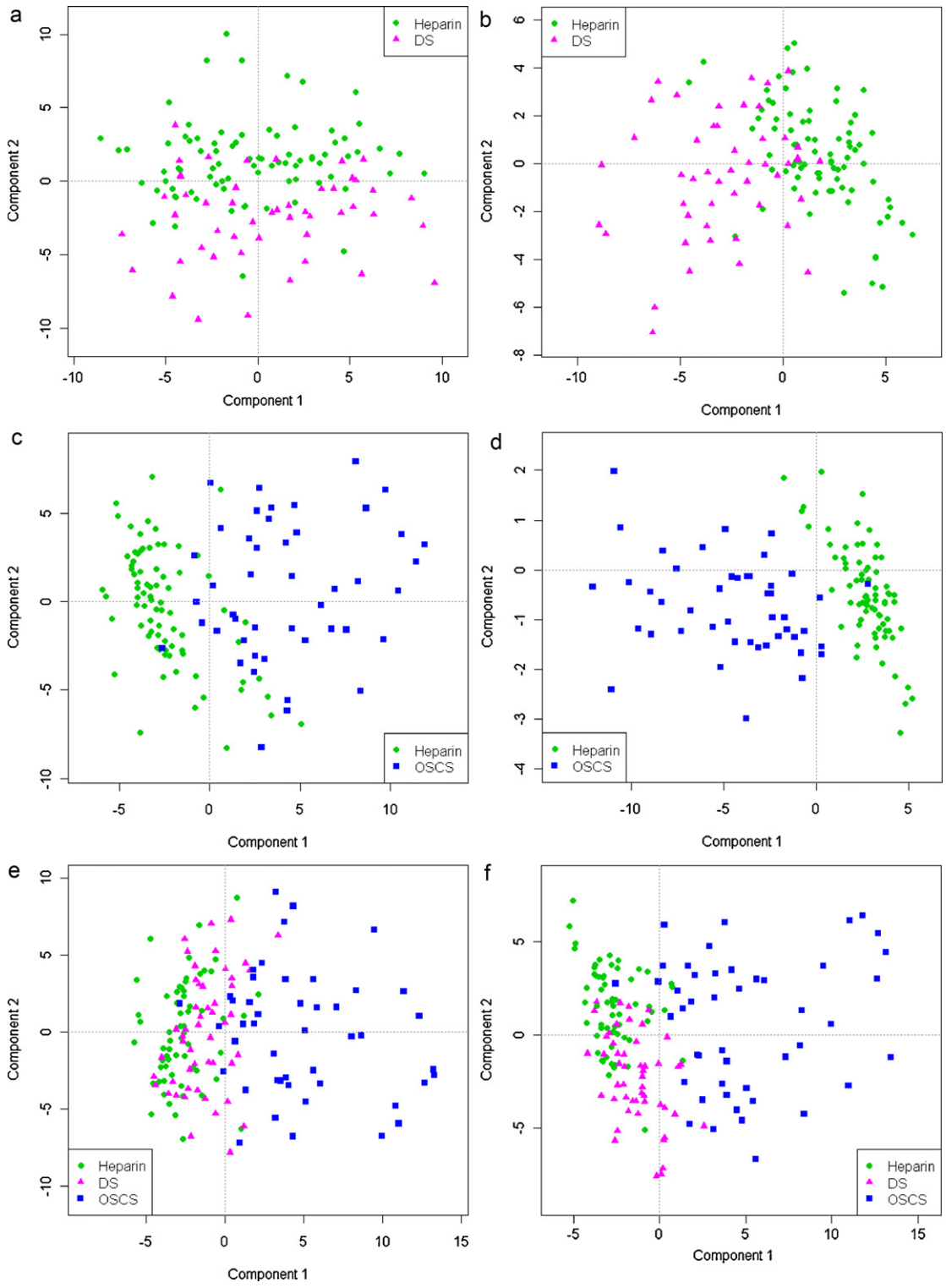

Fig. 2. Scores plots. (a) PCA, *Heparin* vs *DS*; (b) PLS-DA, *Heparin* vs *DS*; (c) PCA, *Heparin* vs *OSCS*; (d) PLS-DA, *Heparin* vs *OSCS*; (e) PCA, *Heparin* vs *DS* vs *OSCS*; (f) PLS-DA, *Heparin* vs *DS* vs *OSCS*.

sification problems. In addition, the PC scores can be used as inputs to multivariate analyses.

The PCA score plots mapped in two dimensions (PC1 vs PC2), obtained from analysis of the ^{1}H NMR spectra for a subset of heparin samples, are shown in Fig. 2a, c and e. Each point on the plots represents one spectrum of an individual sample, and points of the same color indicate samples of the same origin, such as pure heparin (*Heparin*), heparin with the impurity DS (*DS*), or heparin with the contaminant OSCS (*OSCS*). The spectra with similar characteristics form a cluster and the variations along the PC axes maximize the differences between the spectra.

The *Heparin* and *DS* samples were only partially separated into distinct domains using two-dimensional PCA (Fig. 2a). This partial separation was anticipated, in view of the similarities in the NMR spectra of heparin and DS. Furthermore, the intermingled points for the *DS* and *Heparin* samples seen in Fig. 2a correspond to cases

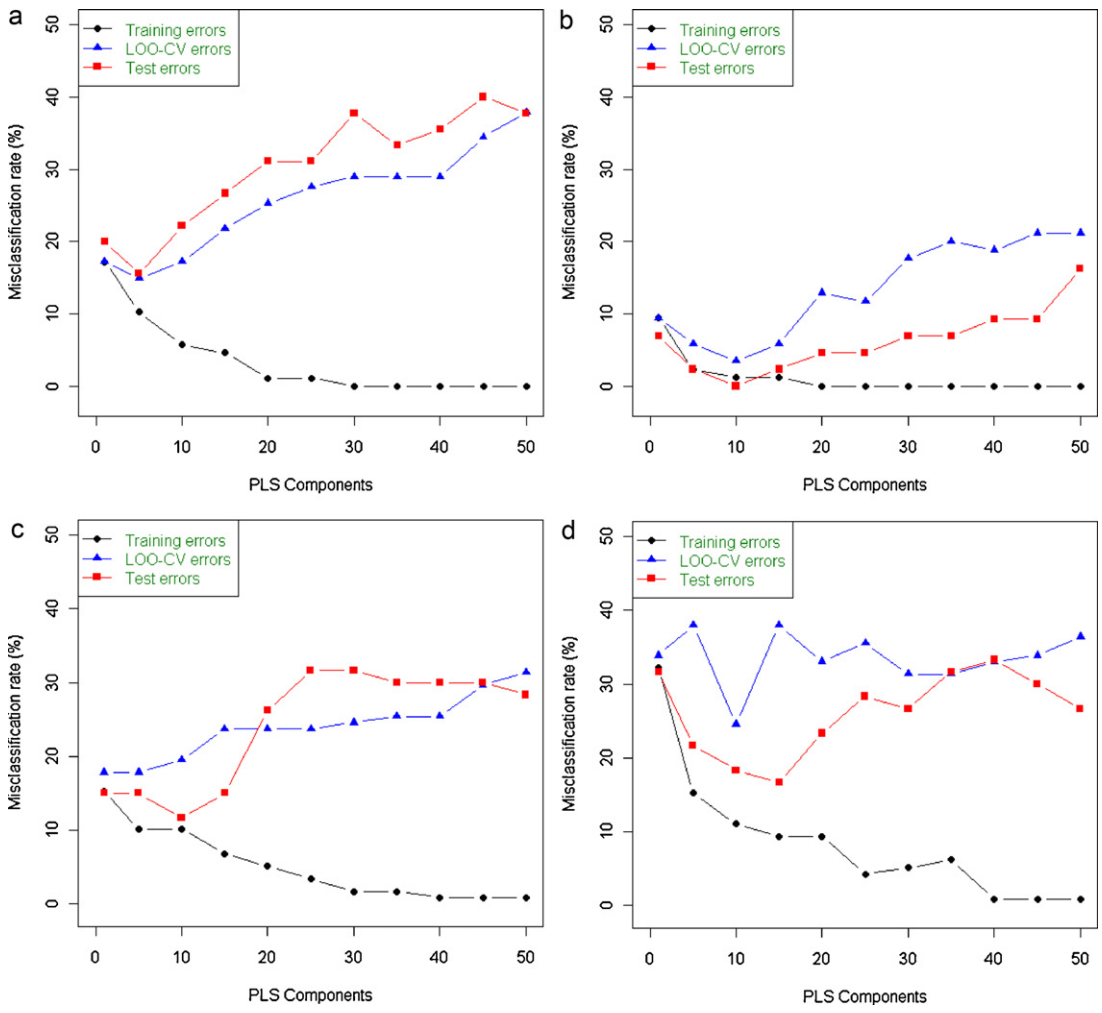

Fig. 3. Misclassification rate as a function of the number of PLS components for the PLS-DA model. (a) *Heparin* vs *DS*; (b) *Heparin* vs *OSCS*; (c) *Heparin* vs [*DS* + *OSCS*]; (d) *Heparin* vs *DS* vs *OSCS*.

in which the DS content is near the 1.0% boundary for impurity acceptance.

With respect to discriminating the *Heparin* and *OSCS* samples, the PC1 vs PC2 scores plot shows clean separation into two distinct clusters (Fig. 2c). The cluster is tighter for the *Heparin* group than for the *OSCS* group, consistent with the smaller variability of DS content in the former relative to OSCS content in the latter.

Discrimination of the *Heparin*, *DS* and *OSCS* samples is apparent in the PC1 vs PC2 scores plot (Fig. 2e). PC1 plays a dominant role in separating the three types of samples into distinct clusters, albeit with some sample overlap. The contribution of PC2 is more in capturing within-sample variability in each case. The extensive spread of the *OSCS* points in the plot reflects the great variability in both OSCS and DS content in these samples. To achieve further separation and quantitatively classify these samples, supervised analysis of the pattern recognition (*i.e.*, PLS-DA, LDA, and kNN) was performed.

3.2.2. Partial least squares discriminant analysis (PLS-DA)

PLS-DA is a linear regression approach in which the multivariate variables from the observations are correlated with the class affiliation of each sample [21]. PLS-DA attempts to build models that can maximize the separation among classes of objects. Since the class affiliation of the objects is included in the regression calculation, PLS-DA is a supervised approach. The regression of the latent variables (T) against a "dummy matrix" (Y) describes the variation according to class affiliation, where Y contains the values of 1 and 0 for each class and comprises as many columns as there are classes. For the training set, an observation is assigned the value of 1 for its class affiliation and 0 for the other classes. The output of PLS-DA regression is a matrix which can be used to classify unknown samples. The prediction result from the PLS-DA model is a numeric value. If the value is close to 1, then the test sample is assigned to the modeled class; if the value is close to 0, then the object is unassigned or assigned to another class.

To optimize separation between pure, impure and contaminated heparin samples, and to build predictive models for class identification, PLS-DA was performed using the classes of *Heparin*, *DS* or *OSCS* as the *y* variables. The two-dimensional scores plots of the first and second latent variables (similar to PCs) are displayed in Fig. 2b, d and f. With PLS-DA, nearly all samples were in distinct classes, and clear discrimination of the *Heparin* samples from the *DS* and *OSCS* samples was observed. Here, the *Heparin* samples appeared in a more compact grouping, while the contaminated (*OSCS*) samples were distributed broadly in the scores plot similar to that in the PCA model. PLS-DA correctly classified these samples into three distinct clusters, as shown in Fig. 2f. This supervised clustering approach gave much improved separation compared with the PCA model, and excellent class discrimination was achieved between the three types of heparin samples.

After PLS data compression, PLS-DA classification models were built and tested while increasing the number of PLS components

Table 1
Number and type of misclassifications (errors) by PLS-DA classification model for test sets.

Components	1	2	4	6	8	10	12	14	16	18	20
Model											
Heparin vs *DS*											
Heparin errors/28 samples	4	2	1	1	2	4	4	5	5	5	6
DS errors/17 samples	5	5	6	6	6	6	7	7	7	8	8
Heparin vs *OSCS*											
Heparin errors/28 samples	0	0	0	0	0	0	0	1	1	1	1
OSCS errors/15 samples	3	2	2	1	1	0	0	0	1	1	1
Heparin vs [*DS* + *OSCS*]											
Heparin errors/28 samples	3	4	2	1	2	2	3	3	4	5	8
[DS + OSCS] errors/32 samples	9	6	7	6	5	5	5	5	5	6	8
Heparin vs *DS* vs *OSCS*											
Heparin errors/28 samples	4	3	1	1	1	2	3	3	3	3	4
DS errors/17 samples	7	7	7	8	8	8	7	7	5	7	8
OSCS errors/15 samples	6	6	5	4	2	1	1	1	1	1	2

starting at 1. The number of correct classifications in both the training and test sets was taken as a measure of performance. Fig. 3 illustrates the evolution of the misclassification rates in the training and test sets as a function of the number of PLS components in the model. As expected for the training set, the number of correct classifications increased with the number of dimensions (PCs). For any model, the misclassification rates were small even with few PLS components and reached a plateau at which all the rates approached zero after 20–40 components.

Leave-one-out cross-validation (LOO-CV) was employed to select the model with the optimal number of PLS components that minimize the misclassification rate. For LOO-CV, the data set was split into s segments: the training was performed on the $(s-1)$ blocks, and the testing was conducted on the objects belonging to the *sth* subset. To predict all the objects, this process was repeated s times through block permutation [29]. Classification rates of 85%, 97% and 82% were obtained for *Heparin* vs *DS*, *Heparin* vs *OSCS*, and *Heparin* vs [*DS* + *OSCS*] models, respectively. In addition, a 75% classification rate was attained by the threefold *Heparin* vs *DS* vs *OSCS* model. The majority of misclassifications between *Heparin* and *DS* involved cases where the DS content was close to the 1.0% DS boundary between the two classes, as measured by SAX-HPLC measurements.

The true test of the model depends on its performance when applied to an external test set of samples that were not employed for building the model. Consequently, the model was validated using an external test set of 60 samples. The results, plotted in Fig. 3, point to the same conclusions as described above for the LOO-CV. By increasing the number of PLS components incrementally, it was observed that the classification rates were optimal for the *Heparin* vs *DS* (84%), *Heparin* vs *OSCS* (100%), and *Heparin* vs [*DS* + *OSCS*] (88%) models when the number of PCs = 2–6, 10–12, and 6–10, respectively. Even for the threefold *Heparin* vs *DS* vs *OSCS* model, the classification performance was 85% using 16 PCs.

The results for the corresponding test sets are presented in Table 1. For the *Heparin* vs *DS* model using 4–6 PCs, misclassification of *Heparin* as *DS* occurred only once and *DS* as *Heparin* six times. In nearly all of these cases the DS content was 1.06–1.20%, *i.e.*, near the 1.0% boundary specifying the two classes.

For the *Heparin* vs *OSCS* model using 1–12 PCs, misclassification of *Heparin* as *OSCS* was zero and *OSCS* for *Heparin* varied from 0 to 3. The number of misclassifications was zero (100% success rate) for the *Heparin* vs *OSCS* model using 10–12 PCs.

For the *Heparin* vs [*DS* + *OSCS*] model using 8–10 PCs, only two *Heparin* samples and five samples in the [*DS* + *OSCS*] group were misclassified. As noted for the *Heparin* vs *DS* model, in most cases these misclassifications occurred when the DS content was near the 1.0% DS boundary defining the *Heparin* and *DS* classes. The same interpretation applies to the threefold *Heparin* vs *DS* vs *OSCS* model, where most of the misclassifications involved samples near the 1.0% DS borderline between *Heparin* and *DS*. Notably, the discrimination between the *Heparin* and *OSCS* samples was 100%.

3.2.3. Linear discriminant analysis (LDA)

As an alternative approach, LDA was employed to classify the *Heparin*, *DS* and *OSCS* samples based on predefined classes. LDA is a well established method for supervised pattern recognition as well as dimension reduction for variable selection [21]. In LDA, a linear function of the dataset is sought so that the ratio of between-class variance is maximized and the ratio of within-class variance is minimized, and finally the optimal separation among the given classes is achieved. Like PLS-DA, the ultimate aim of LDA is to qualitatively predict the group affiliation for unknown samples. Discrimination of the classes is performed by calculating the Mahalanobis distance of a sample from the center of gravity of each specified class, then assigning the sample to the class associated with the smallest distance. A test sample was correctly classified if it was located nearest to the center of gravity of its actual class. Otherwise, the sample would be (incorrectly) classified to another class for which the Mahalanobis distance was the smallest.

In order to select a subset of the original variables that affords the maximum improvement of the discriminating ability between classes, stepwise linear discriminant analysis (SLDA) was performed before LDA analysis. SLDA employs an aggregative procedure, which starts with no variables in the model and adds the variables with the greatest discriminating ability in successive steps [30]. In SLDA, Wilks' lambda is employed as a selection criterion to determine the variables included in the procedure. Wilks' lambda is defined as the ratio of the intra-class covariance to the total covariance; hence, its value varies between 0 and 1. A value close to 0 denotes that the classes are well separated, while a value close to 1 denotes that the classes are poorly separated.

As the first step, the variable that best discriminates the groups is selected for the model. Each successive step involves evaluation of all remaining variables in order to select the one that yields the minimum intra-class covariance, *i.e.*, the smallest Wilks' lambda, which implies that the within-class sum of squares is minimized while the inter-class sum of squares is maximized. The selection procedure stops when all variables have been evaluated. Preliminary variable reduction using SLDA led to the selection of a series of variables from 1 to 20 (Table 2).

After variable selection by dimension reduction, LDA analysis was conducted using the squared Mahalanobis distance from the centers of gravity of each group for assigning the class affiliation of each sample. The results show that class discrimination improved markedly after variable selection (Table 3). For the training set, the

Table 2
The variables (ppm) selected from stepwise linear discriminant analysis (SLDA) for various models.

Order	*Heparin* vs *DS*	*Heparin* vs *OSCS*	*Heparin* vs [*DS* + *OSCS*]	*Heparin* vs *DS* vs *OSCS*
1	2.07	2.16	2.07	2.10
2	3.61	2.07	4.49	3.86
3	5.34	4.49	2.16	3.52
4	2.16	4.16	4.16	4.49
5	2.13	4.04	4.46	5.16
6	4.61	3.55	5.16	3.58
7	2.10	4.52	5.10	2.16
8	3.95	3.64	5.61	3.95
9	5.67	5.61	4.28	4.46
10	4.04	5.67	3.55	5.00
11	5.43	4.37	4.94	4.43
12	3.70	5.25	5.49	3.70
13	4.46	3.73	4.97	5.13
14	3.76	5.03	4.61	5.03
15	3.73	2.13	4.22	5.46
16	5.40	5.49	5.19	4.64
17	3.67	3.67	5.43	4.13
18	4.01	4.10	4.34	4.16
19	5.19	5.28	5.58	4.28
20	5.31	5.19	5.25	4.22

success rates approached 100% with increasing the number of variables. The *Heparin* vs *OSCS* model required very few variables to achieve 100% success rates due to the clear distinction in spectral features between heparin and OSCS. Cross validation and external validation studies indicated that model performance reached a maximum using an intermediate number of variables. LDA models typically include a set of tunable parameters, the number of which increases with the number of variables. While even models with complex relationships in the sample can usually be fit quite well by using enough tunable parameters, this typically leads to much higher error rates for the test set than for the training set as occurred in the present instance.

The risks of over-fitting can be alleviated by selecting the optimal number of variables, which was determined by the successful rate of classifications using LOO-CV and validation with external test sets. Optimal success rates, varying from 89% to 100%, for the *Heparin* vs *DS*, *Heparin* vs *OSCS*, *Heparin* vs [*DS* + *OSCS*] models were achieved using only 6–14 variables depending on the specific model and testing procedure (Table 3). In the same way, the threefold *Heparin* vs *DS* vs *OSCS* model achieved an optimal success rate of 90% using 10–12 variables. Once again, the majority of misclassifications are attributed to *Heparin* and *DS* samples in which the DS content was near the 1.0% boundary between the two classes.

With respect to classification of individual samples and overall success rates, the performance of LDA was comparable to PLS-DA for the *Heparin* vs *OSCS* model and superior to PLS-DA for other three models. For the external test set under optimal conditions, the success rates for the *Heparin* vs *DS*, *Heparin* vs [*DS* + *OSCS*], and *Heparin* vs *DS* vs *OSCS* models were respectively 89%, 93%, and 90% using LDA compared to 84%, 88% and 85% using PLS-DA.

3.2.4. *k*-Nearest-neighbor (kNN)

The kNN method calculates the distances between a new object (a test data point) and all objects in the training set in *p*-dimensional variable space [31–33]. Unlike PLS-DA and LDA, the kNN approach avoids the need for model generation. Neighbor determination is calculated by the Euclidean distance, and the nearest *k* objects are used to estimate the class affiliation of the test object. By applying the majority rule, the new object is assigned to the class of the majority of the *k* objects, *i.e.*, the prediction is related to a majority vote among the neighbors. To correctly assign the group affiliation for a test data point, this technique requires tuning of the adjustable parameter *k* (*i.e.*, the optimal number of nearest neighbors). Values of *k* that are too small (leading to underfitting) or too large (leading to overfitting) can lead to poor classification of new objects. By testing a series of *k* values and assessing the prediction performance, the optimal value of *k* corresponds to that giving lowest number of misclassifications.

Table 3
Performance of LDA classification models under different variables selected from SLDA.

Number of variables		2	4	6	8	10	12	14	16	18	20
Heparin vs *DS*											
Training set	Errors/87 samples	14	12	10	10	9	9	8	6	5	3
	Success rates (%)	84	86	89	89	90	90	91	93	94	97
CV set	Errors/87 samples	15	13	12	12	10	10	12	13	14	14
	Success rates (%)	83	85	86	86	89	89	86	85	84	84
Test set	Errors/45 samples	7	6	5	5	6	6	7	8	8	10
	Success rates (%)	84	87	89	89	87	87	84	82	82	78
Heparin vs *OSCS*											
Training set	Errors/85 samples	6	4	4	2	1	1	0	0	0	0
	Success rates (%)	93	95	95	98	99	99	100	100	100	100
CV set	Errors/85 samples	6	5	4	4	2	0	1	2	3	5
	Success rates (%)	93	94	95	95	98	100	99	98	97	94
Test set	Errors/43 samples	2	1	1	1	0	0	1	2	2	3
	Success rates (%)	95	98	98	98	100	100	98	95	95	93
Heparin vs [*DS* + *OSCS*]											
Training set	Errors/118 samples	17	15	14	14	13	13	12	10	9	9
	Success rates (%)	86	87	88	88	89	89	90	92	93	93
CV set	Errors/118 samples	19	18	18	16	14	11	10	12	15	17
	Success rates (%)	84	85	85	86	88	91	92	90	87	86
Test set	Errors/60 samples	7	6	5	5	4	5	6	6	6	8
	Success rates (%)	88	90	92	92	93	92	90	90	90	87
Heparin vs *DS* vs *OSCS*											
Training set	Errors/118 samples	26	24	21	19	16	14	12	12	10	8
	Success rates (%)	78	80	82	84	86	88	90	90	92	93
CV set	Errors/118 samples	28	27	25	19	15	13	16	18	19	21
	Success rates (%)	76	77	79	84	87	89	86	85	84	82
Test set	Errors/60 samples	12	11	10	9	6	6	8	8	10	10
	Success rates	80	82	83	85	90	90	87	87	83	83

Table 4
Performance of kNN classification models for the original data.

	Model	*Heparin* vs *DS*	*Heparin* vs *OSCS*	*Heparin* vs [*DS* + *OSCS*]	*Heparin* vs *DS* vs *OSCS*
$k = 3$	Training set				
	Errors/samples	7/87	1/85	13/118	16/118
	Success rate (%)	92	99	89	86
	LOO-CV set				
	Errors/samples	16/87	4/85	25/118	32/118
	Success rate (%)	82	95	79	73
	Test set				
	Errors/samples	14/45	4/43	11/60	19/60
	Success rate (%)	69	91	82	68
$k = 5$	Training set				
	Errors/samples	12/87	2/85	17/118	21/118
	Success rate (%)	86	98	86	82
	LOO-CV set				
	Errors/samples	17/87	5/85	25/118	30/118
	Success rate (%)	81	94	79	75
	Test set				
	Errors/samples	13/45	4/43	11/60	22/60
	Success rate (%)	71	91	82	63
$k = 7$	Training set				
	Errors/samples	13/87	2/85	17/118	20/118
	Success rate (%)	85	98	86	83
	LOO-CV set				
	Errors/samples	14/87	5/85	27/118	33/118
	Success rate (%)	84	94	77	72
	Test set				
	Errors/samples	13/45	4/43	13/60	21/60
	Success rate (%)	71	91	78	65

The kNN method was implemented to evaluate its performance for classification. Various k values (3, 5 or 7) were tested using the all-variable data set, and the success rates for the training set, LOO-CV, and the test set are summarized in Table 4. Overall, the results obtained were inferior for kNN compared with LDA and PLS-DA. For example, the success rates for the *Heparin* vs *DS*, *Heparin* vs *OSCS*, *Heparin* vs [*DS* + *OSCS*], and *Heparin* vs *DS* vs *OSCS* models using $k = 3$ were respectively 69%, 91%, 82% and 68% for the test set.

To obtain better classification results, the PCA scores were employed as inputs to build the kNN models. Various combinations of PCs and k values were investigated, and the results are summarized in Table 5. Unlike the PLS-DA and LDA models where the misclassification rates for the training set decreased monotonically to 0% as the number of PCs or variables increased, the misclassification rates of the kNN models for the training set fluctuated within a range of values. The optimal performance of the kNN

Table 5
Performance of PCA-kNN classification models under different PCs.

PCs		5	10	15	20	25	30	35	40	45	50	55	60
Heparin vs *DS* ($k = 2$)													
Training set	Errors/87 samples	13	11	7	5	12	8	10	12	10	13	15	14
	Success rates (%)	85	87	92	94	86	91	89	86	89	85	83	84
CV set	Errors/87 samples	25	20	17	20	25	25	27	22	29	34	31	33
	Success rates (%)	71	77	80	77	71	71	69	75	67	61	64	62
Test set	Errors/45 samples	12	15	16	12	10	14	12	15	15	12	16	19
	Success rates (%)	73	67	64	73	78	69	73	67	67	73	64	58
Heparin vs *OSCS* ($k = 4$)													
Training set	Errors/85 samples	6	3	5	5	9	8	8	11	11	16	13	19
	Success rates (%)	93	96	94	94	89	91	91	87	87	81	85	78
CV set	Errors/85 samples	10	13	11	10	14	18	19	25	22	24	25	26
	Success rates (%)	88	85	87	88	84	79	78	71	74	72	71	69
Test set	Errors/43 samples	37	38	40	39	39	37	30	33	33	31	30	33
	Success rates (%)	86	88	93	91	91	86	70	77	77	72	70	77
Heparin vs [*DS* + *OSCS*] ($k = 3$)													
Training set	Errors/118 samples	17	10	13	17	19	11	16	14	18	17	19	25
	Success rates (%)	86	92	89	86	84	91	86	88	85	86	84	79
CV set	Errors/118 samples	23	30	26	34	33	39	31	28	34	36	34	43
	Success rates (%)	81	75	78	71	72	67	74	76	71	69	71	64
Test set	Errors/60 samples	13	13	12	9	17	15	17	19	23	22	22	21
	Success rates (%)	78	78	80	85	72	75	72	68	62	63	63	65
Heparin vs *DS* vs *OSCS* ($k = 3$)													
Training set	Errors/118 samples	18	13	19	23	22	17	21	21	23	23	25	32
	Success rates (%)	85	89	84	81	81	86	82	82	81	81	78	73
CV set	Errors/118 samples	30	39	32	42	42	40	43	43	47	41	46	52
	Success rates (%)	75	67	73	64	64	66	64	64	60	65	61	56
Test set	Errors/60 samples	21	19	18	15	20	23	23	23	25	24	27	27
	Success rates	65	68	70	75	67	62	62	62	58	60	55	55

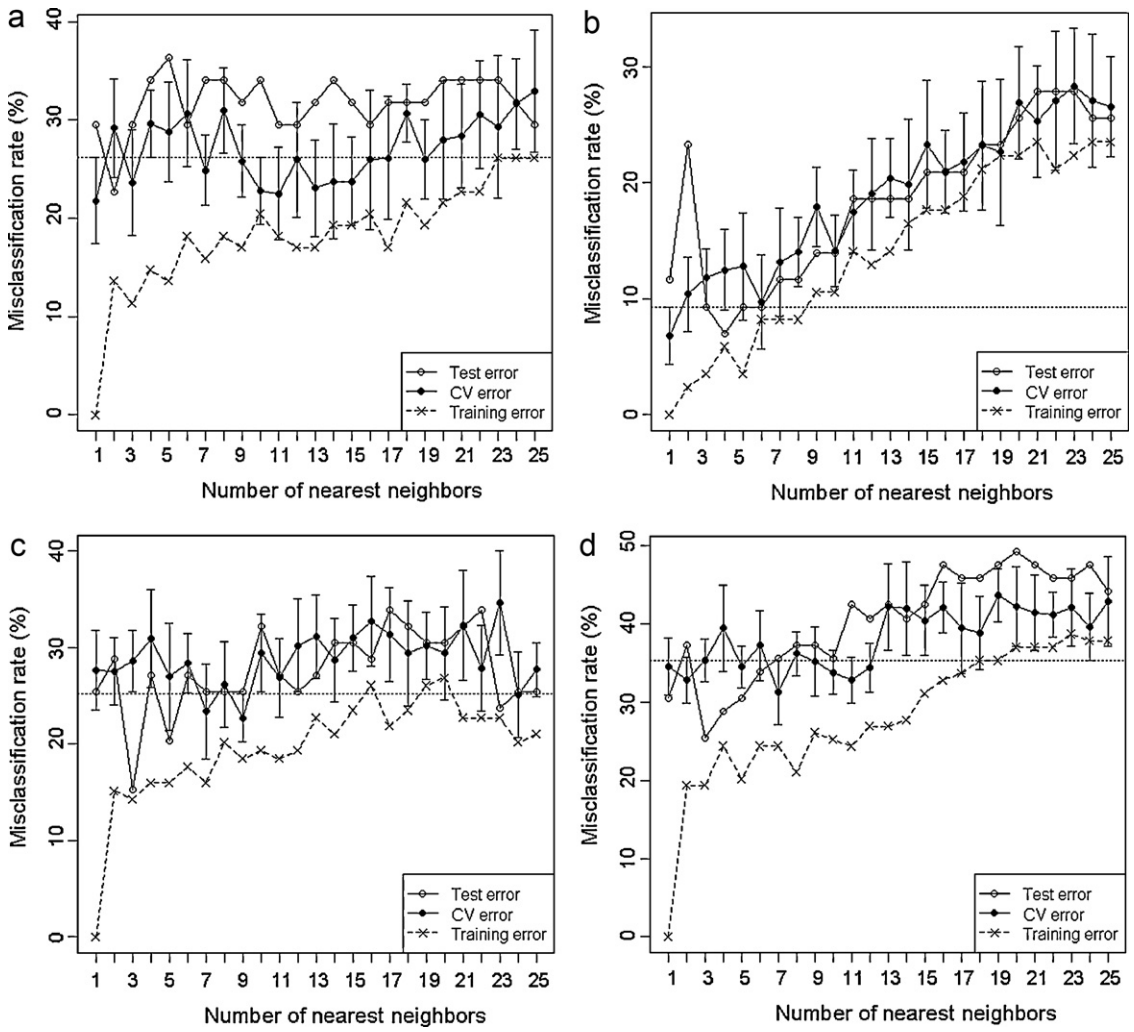

Fig. 4. kNN classification for heparin-contaminant data over the range $k = 1$ to $k = 25$. (a) *Heparin* vs *DS* (PCs = 25); (b) *Heparin* vs *OSCS* (PCs = 15); (c) *Heparin* vs [*DS* + *OSCS*] (PCs = 20); (d) *Heparin* vs *DS* vs *OSCS* (PCs = 20).

model was achieved for the *Heparin* vs *DS*, *Heparin* vs *OSCS*, *Heparin* vs [*DS* + *OSCS*], and *Heparin* vs *DS* vs *OSCS* models using 15–25 PCs depending on the specific model.

The misclassification rates for nearest neighbors *k* from 1 to 25 are plotted in Fig. 4. The black dots and the vertical bars represent the means as well as mean ± 1 standard error for the misclassification rates using LOO-CV [21]. The smallest LOO-CV error is depicted by a dotted horizontal line corresponding to the position of the mean plus one standard error. For the training sets, the misclassification rate was always zero for $k = 1$ and increased with larger *k* values for all four models. The test sets showed a similar pattern, *i.e.*, the misclassification rates varied within a tight range, except the *Heparin* vs *OSCS* model for which the rates rose for $k > 4$. The optimal *k* values of 2, 4, 3 and 3 respectively were for the *Heparin* vs *DS*, *Heparin* vs *OSCS*, *Heparin* vs [*DS* + *OSCS*], and *Heparin* vs *DS* vs *OSCS* models.

When the predictive ability was evaluated for the external test set based on the above analysis for different numbers of PCs and a series of *k* values, the optimal success rates were 78%, 93%, 83% and 75% for the four models as shown in Table 5. The classification performance was inferior for kNN compared with PLS-DS and LDA. For the *Heparin* vs *DS* model, one heparin sample was misclassified as *DS* but nine out of the seventeen *DS* test samples were misclassified as *Heparin*. Unlike PLS-DA and LDA, kNN was unable to completely discriminate *Heparin* and *OSCS*. For the *Heparin* vs [*DS* + *OSCS*] model, three *Heparin* samples were misclassified as [*DS* + *OSCS*] while six *DS* samples and one *OSCS* sample were misclassified as *Heparin*. Likewise for the threefold *Heparin* vs *DS* vs *OSCS* model, kNN produced a total of fifteen misclassifications.

4. Conclusions

In the present study, we applied multivariate chemometric approaches in combination with ^{1}H NMR spectroscopy for qualitative and quantitative analysis of heparin samples that may possess dermatan sulfate (DS) impurities, oversulfated chondroitin sulfate (OSCS) contaminants, or both. We show that these chemometric methods (PCA, PLS-DA, LDA or kNN) are useful tools for the exploration and visualization of heparin NMR spectral data, and for the generation of classification models with outstanding performance attributes. The large number of original variables (74) was reduced by chemometric methods into a much smaller number of new variables (PCs, or latent variables) for effective clustering and classification. The degree of success of the classification models in discriminating the samples of pure heparin (*Heparin*) from those containing the impurity DS (*DS*) and the contaminant OSCS (*OSCS*) depended on the specific chemometric procedures for choosing the appropriate variables.

The well-known unsupervised chemometric method of PCA was used to explore the similarities and differences in the complex pattern of overlapping ^{1}H NMR signals found in the heparin spectra. The PCA results showed that the samples were separated into two

distinct clusters for the *Heparin* vs *OSCS* groups, but the distinction between *Heparin* and *DS* was less evident. Excellent discrimination of the *Heparin* samples from those samples containing impurities (*DS*) and contaminants (*OSCS*) was achieved with the supervised method PLS-DA.

The predictive performance of the models obtained from PLS-DA and LDA were outstanding in differentiating *Heparin* from *DS* and *OSCS* with very few misclassifications. In all cases, better classification rates (fewer misclassifications) were attained for *Heparin* vs *OSCS* models than for *Heparin* vs *DS* models regardless of the clustering and classification approach. Under optimal conditions, success rates of 100% were frequently achieved for discrimination between *Heparin* and *OSCS* samples. This outcome is plausible, in view of the much closer similarity in the ^{1}H NMR spectral patterns between *Heparin* and *DS* than between *Heparin* and *OSCS*. The LDA approach outperformed PLS-DA (89% vs 84%) for discrimination of the *Heparin* and *DS* samples.

In summary, the present study reveals that ^{1}H NMR spectroscopy, in combination with multivariate chemometric methods such as PLS-DA and LDA, represent an effective strategy for fast and reliable identification of impurities (DS) and contaminants (OSCS) in heparin API samples. The pattern recognition approach applied here may be useful in monitoring purity of other complex naturally derived compounds.

References

[1] A.W. McMahon, R.G. Pratt, T.A. Hammad, S. Kozlowski, E. Zhou, S. Lu, C.G. Kulick, T. Mallick, G.D. Pan, Description of hypersensitivity adverse events following administration of heparin that was potentially contaminated with oversulfated chondroitin sulfate in early 2008, Pharmacoepidemiol. Drug Safety 19 (2010) 921–933.

[2] S. Beni, J.F.K. Limtiaco, C.K. Larive, Analysis and characterization of heparin impurities, Anal. Bioanal. Chem., in press [E-Pub 2010 Sep3].

[3] C. Tami, M. Puig, J.C. Reepmeyer, H. Ye, D.A. D'Avignon, L. Buhse, D. Verthelyi, Inhibition of Taq polymerase as a method for screening heparin for oversulfated contaminants, Biomaterials 29 (2008) 4808–4814.

[4] J.A. Spencer, J.F. Kauffman, J.C. Reepmeyer, C.M. Gryniewicz, W. Ye, D.Y. Toler, L.F. Buhse, B.J. Westenberger, Screening of heparin API by near infrared reflectance and Raman spectroscopy, J. Pharm. Sci. 98 (2009) 3540–3547.

[5] M.L. Trehy, J.C. Reepmeyer, R.E. Kolinski, B.J. Westenberger, L.F. Buhse, Analysis of heparin sodium by SAX/HPLC for contaminants and impurities, J. Pharm. Biomed. Anal. 49 (2009) 670–673.

[6] T. Wielgos, K. Havel, N. Ivanova, R. Weinberger, Determination of impurities in heparin by capillary electrophoresis using high molarity phosphate buffers, J. Pharm. Biomed. Anal. 49 (2009) 319–326.

[7] M. Guerrini, D. Beccati, Z. Shriver, A. Naggi, K. Viswanathan, A. Bisio, I. Capila, J.C. Lansing, S. Guglieri, B. Fraser, A. Al-Hakim, N.S. Gunay, Z. Zhang, L. Robinson, L. Buhse, M. Nasr, J. Woodcock, R. Langer, G. Venkataraman, R.J. Linhardt, B. Casu, G. Torri, R. Sasisekharan, Oversulfated chondroitin sulfate is a contaminant in heparin associated with adverse clinical events, Nat. Biotechnol. 26 (2008) 669–675.

[8] D.L. Rabenstein, Heparin and heparan sulfate: structure and function, Nat. Prod. Rep. 19 (2002) 312–331.

[9] M. Guerrini, Z. Zhang, Z. Shriver, A. Naggi, S. Masuko, R. Langer, B. Casu, R.J. Linhardt, G. Torri, R. Sasisekharan, Orthogonal analytical approaches to detect potential contaminants in heparin, PNAS 106 (2009) 16956–16961.

[10] T.K. Kishimoto, K. Viswanathan, T. Ganguly, S. Elankumaran, S. Smith, K. Pelzer, J.C. Lansing, N. Sriranganathan, G. Zhao, Z. Galcheva-Gargova, A. Al-Hakim, G.S. Bailey, B. Fraser, S. Roy, T. Rogers-Cotrone, L. Buhse, M. Whary, J. Fox, M. Nasr, G.J.D. Pan, Z. Shriver, R.S. Langer, G. Venkataraman, K.F. Austen, J. Woodcock, R. Sasisekharan, Contaminated heparin associated with adverse clinical events and activation of the contact system, N. Engl. J. Med. 358 (2008) 2457–2467.

[11] T. Beyer, B. Diehl, G. Randel, E. Humpfer, H. Schäfer, M. Spraul, C. Schollmayer, U. Holzgrabe, Quality assessment of unfractionated heparin using ^{1}H nuclear magnetic resonance spectroscopy, J. Pharm. Biomed. Anal. 48 (2008) 13–19.

[12] R. Domanig, W. Jöbstl, S. Gruber, T. Freudemann, One-dimensional cellulose acetate plate electrophoresis—a feasible method for analysis of dermatan sulfate and other glycosaminoglycan impurities in pharmaceutical heparin, J. Pharm. Biomed. Anal. 49 (2009) 151–155.

[13] D.A. Keire, H. Ye, M.L. Trehy, W. Ye, R.E. Kolinski, B.J. Westenberger, L.F. Buhse, M. Nasr, A. Al-Hakim, Characterization of currently marketed heparin products: key tests for quality assurance, Anal. Bioanal. Chem., in press [E-Pub 2010 Aug1].

[14] P. Bigler, R. Brenneisen, Improved impurity fingerprinting of heparin by high resolution ^{1}H NMR Spectroscopy, J. Pharm. Biomed. Anal. 49 (2009) 1060–1064.

[15] J.T. King, U.R. Desai, A capillary electrophoretic method for fingerprinting low molecular weight heparins, Anal. Biochem. 380 (2008) 229–234.

[16] D.A. Keire, M.L. Trehy, J.C. Reepmeyer, R.E. Kolinski, J. Dunn, W. Ye, B.J. Westenberger, L.F. Buhse, Analysis of crude heparin by ^{1}H NMR, capillary electrophoresis, and strong-anion-exchange-HPLC for contamination by over sulfated chondroitin sulfate, J. Pharm. Biomed. Anal. 51 (2010) 921–926.

[17] D.A. Keire, D.J. Mans, H. Ye, R.E. Kolinski, L.F. Buhse, Assay of possible economically motivated additives or native impurities levels in heparin by ^{1}H NMR, SAX-HPLC, and anticoagulation time approaches, J. Pharm. Biomed. Anal. 52 (2010) 656–664.

[18] Q. Zang, D.A. Keire, R.D. Wood, L.F. Buhse, C.M.V. Moore, M. Nasr, A. Al-Hakim, M.L. Trehy, W.J. Welsh, Determination of galactosamine in heparin by ^{1}H NMR spectroscopy coupled with variable selection and multivariate regression analysis, Anal. Bioanal. Chem., in press [E-Pub 2010 Oct 16].

[19] U.S. Pharmacopeia, Pharmacopeial Forum, March–April 35 (2009) p. 257.

[20] R Development Core Team, R: A Language and Environment for Statistical Computing, 2009.

[21] K. Varmuza, P. Filzmoser, Introduction to Multivariate Statistical Analysis in Chemometrics, CRC Press, Boca Raton, FL, USA, 2009.

[22] J. Maindonald, J. Braun, Data Analysis and Graphics using R, Cambridge University Press, Cambridge, UK, 2003.

[23] M. Forina, S. Lanteri, C. Armanino, C. Casolino, M. Casale, V-Parvus 2007, http://www.parvus.unige.it.

[24] V. Ruiz-Calero, J. Saurina, M.T. Galceran, S. Hernández-Cassou, L. Puignou, Estimation of the composition of heparin mixtures from various origins using proton nuclear magnetic resonance and multivariate calibration methods, Anal. Bioanal. Chem. 373 (2002) 259–265.

[25] V. Ruiz-Calero, J. Saurina, M.T. Galceran, S. Hernández-Cassou, L. Puignou, Potentiality of proton nuclear magnetic resonance and multivariate calibration methods for the determination of dermatan sulfate contamination in heparin samples, Analyst 125 (2000) 933–938.

[26] V. Ruiz-Calero, J. Saurina, S. Hernández-Cassou, M.T. Galceran, L. Puignou, Proton nuclear magnetic resonance characterization of glycosaminolgycans using chemometric techniques, Analyst 127 (2002) 407–415.

[27] T.R. Rudd, M.A. Skidmore, S.E. Guimond, C. Cosentino, G. Torri, D.G. Fernig, R.M. Lauder, M. Guerrini, E.A. Yates, Glycosaminoglycan origin and structure revealed by multivariate analysis of NMR and CD spectra, Glycobiology 19 (2009) 52–67.

[28] M. Foot, M. Mulholland, Classification of chondroitin sulfate A, chondroitin sulfate C, glucosamine hydrochloride and glucosamine 6 sulfate using chemometric techniques, J. Pharm. Biomed. Anal. 38 (2005) 397–407.

[29] E.K. Kemsley, Discriminant analysis of high-dimensional data: a comparison of principal components analysis and partial least squares data reduction methods, Chemom. Intell. Lab. Syst. 33 (1996) 47–61.

[30] M. Forina, P. Oliveri, S. Lanteri, M. Casale, Class-modeling techniques, classic and new, for old and new problems, Chemom. Intell. Lab. Syst. 93 (2008) 132–148.

[31] W.J. Welsh, W. Lin, S.H. Tersigni, E. Collantes, R. Duta, M.S. Carey, W.L. Zielinski, J. Brower, J.A. Spencer, T.P. Layloff, Pharmaceutical fingerprinting: evaluation of neural networks and chemometric techniques for distinguishing among same-product manufacturers, Anal. Chem. 68 (1996) 3473–3482.

[32] I.V. Tetko, A.E.P. Villa, T.I. Aksenova, W.L. Zielinski, J. Brower, E.R. Collantes, W.J. Welsh, Application of a pruning algorithm to optimize artificial neural networks for pharmaceutical fingerprinting, J. Chem. Inform. Comput. Sci. 38 (1998) 660–668.

[33] E.R. Collantes, R. Duta, W.J. Welsh, W.L. Zielinski, J. Brower, Preprocessing of HPLC trace impurity patterns by wavelet packets for pharmaceutical fingerprinting using artificial neural networks, Anal. Chem. 69 (1997) 1392–1397.