NMR spectroscopy and chemometric models to detect a specific non-porcine ruminant contaminant in pharmaceutical heparin.

Erika Colombo[a]*, Lucio Mauri[a], Maria Marinozzi[a], Timothy R. Rudd[b,c], Edwin A. Yates[c], Davide Ballabio[d#], Marco Guerrini[a#]

[a] Institute for Chemical and Biochemical Research G. Ronzoni, via G. Colombo 81, 20133 Milan, Italy

[b] National Institute for Biological Standards and Control (NIBSC), Blanche Lane, South Mimms, Potters Bar, Hertfordshire, UK

[c] Department of Biochemistry and Systems Biology, ISMIB, University of Liverpool, Liverpool, L69 7ZB United Kingdom

[d] Milano Chemometrics and QSAR Research Group, Department of Earth and Environmental Sciences, University of Milano Bicocca, P.zza della Scienza, -20126 Milano, Italy

*Present address: Istituto di Ricerche Farmacologiche Mario Negri, Via Mario Negri 2, 20156 Milan, Italy

Corresponding authors: Marco Guerrini, Institute for Chemical and Biochemical Research G. Ronzoni guerrini@ronzoni.it; Davide Ballabio, University of Milano Bicocca davide.ballabio@unimib.it

Keywords

Heparin; heparin origin; NMR; chemometric; multivariate classification

Abstract

Heparin has been used successfully as a clinical antithrombotic for almost one century. Its isolation from animal sources (mostly porcine intestinal mucosa) involves multistep purification processes starting from the slaughterhouse (as mucosa) to the pharmaceutical plant (as the API). This complex supply chain increases the risk of contamination and adulteration, mainly with non-porcine ruminant material. The structural similarity of heparins from different origins, the natural variability of the heparin within samples from each source as well as the structural changes induced by manufacturing processes, require increasingly sophisticated methods capable of detecting low levels of contamination. The application of suitable multivariate classification approaches on API ^{1}H-NMR spectra serve as rapid and reliable tools for product authentication and the detection of contaminants. Soft Independent Modelling of Class Analogies (SIMCA), Discriminant Analysis (DA), Partial Least Square Discriminant Analysis (PLS-DA) and local classification methods (kNN, BNN and N3) were tested on about one hundred certified heparin samples produced by 14 different manufacturers revealing that Partial Least Squares Discriminant Analysis (PLS-DA) provided the best discrimination of contaminated batches, with a balanced accuracy of 97%.

Keywords

Heparin, contamination, classification, NMR

1. Introduction

Heparin is an anticoagulant drug of animal origin listed on the World Health Organization's (WHO) list of essential medicines (https://www.who.int/publications/i/item/WHO-MHP-HPS-EML-2021.02). Heparin exerts its action through interactions with several proteins of the blood clotting cascade, mainly antithrombin (AT), thereby increasing their inhibitory effects [1]. Pharmaceutical heparin, a complex mixture of related polysaccharides, is obtained primarily from porcine and, less frequently, bovine intestinal mucosa by multistep purification processes, which include scraping of the mucosa from the intestine, the extraction of glycosaminoglycans (GAGs) from the mucosa (to generate crude heparin) and subsequent purification to provide the active pharmaceutical ingredient (API) [2]. Since the supply chain for heparin is complex and global, monitoring the entire process (from slaughterhouse to API) is often unfeasible and the risk of contamination or adulteration is consequently high. Whereas the purification of crude heparin to sodium heparin API is always conducted under GMP conditions, not all manufacturers can provide full traceability back to the individual animals. Many companies purchase crude heparin from third parties, who have themselves collected mucosa at numerous small slaughterhouses. These multiple sites, and the attendant lack of traceability of the material obtained from them, only serve to increase the risk of accidental or intentional contamination further.

Although the implementation of orthogonal analytical methods, introduced in pharmacopoeias as a result of heparin adulteration with an unnatural substance in 2007 and 2008, has also reduced the risk of adulteration or contamination with known materials, the possibility of more sophisticated adulteration cannot be excluded [3, 4]. The adulteration events of 2007-2008 were related to shortages of porcine material arising from blue-ear pig disease, endemic in Asia in 2007. Currently, we are witnessing a similar situation with African swine fever afflicting Chinese pigs, the origin of more than half of global heparin [5]. Despite the loss of about 25% of the pig population, this has not affected manufacturing or the distribution of heparin in western countries, at least for the time being, but it will likely result in a shortage of the porcine material which is used for the preparation of heparin. In the Guidance Documents of June 2013 [6], FDA alerted industries to the potential risk of heparin contamination with non-porcine ruminant material. The guidance states that "the control of the animal origin of heparin is important to ensure the safety of drugs and devices that contain heparin and to protect public health" and suggests the application of physico-chemical, immunological, or polymerase chain reaction (PCR)-based methods for the detection of ruminant material in crude heparin. Regarding PCR analysis, concerns were raised about the possibility of contamination of the porcine crude material with purified ruminant material or, with crude ruminant

material that had been subjected to oxidation processes, which could modify the structure of residual DNA/RNA fragments and hence evade detection by PCR.

Recent studies proposed a range of physicochemical methods that can be used to detect contaminants in heparin [7-9]. However, it is much more difficult to detect contamination with non-porcine ruminant material. Heparin APIs from distinct organ/animal sources differ from porcine derived mucosa principally in their degree of sulfation and acetylation and, to some extent, their distinct chain sequences in which sulfated and less sulfated domains are located. For instance, porcine mucosal heparin (PMH) contains more 6-O sulfated residues than bovine mucosal heparin (BMH), but fewer than ovine mucosal heparin (OMH). In addition, the extent of N-acetylation is higher in PMH than in BMH and OMH [10-12]. Moreover, the amount of other minor sequences varies among the different heparin sources. For example, significant 2-O-sulfation of glucuronic acid was found in BMH compared to other sources [10, 13].

To compound the issue further, each heparin type has its own "inherent variability", which depends on the variation in biosynthesis and on structural changes induced by the manufacturing processes. This intrinsic heterogeneity, although still allowing differentiation and identification of each heparin specie, strongly affects the possibility of detecting non-porcine ruminant material in porcine heparin with adequate accuracy [14].

Following the "*heparin crisis*" of 2007-2008, many efforts have been made to characterize heparin [3]. Nuclear magnetic resonance (NMR) spectroscopy emerged as the leading technique in the characterization of GAGs, as it does not require prior separation of the constituent components and provides a broad range of information ranging from a chemical fingerprint (both 1D and 2D-NMR) to structural constraints (quantitative HSQC analysis) [15]]. In addition, the highly reproducible spectra allow analysis of a large number of samples, thereby creating spectral libraries encompassing the natural variability of the product. The limitation of the manual analysis of these large spectral databases is the ability to differentiate samples of interest when comparing complex 1D or 2D spectra. The application of chemometric techniques allows reduction of the complexity of these large data sets and enables definition of features which better define the differences between samples. Various taught chemometric methods addressing the analysis of spectroscopic data have been successfully explored for the detection of contaminants in heparin APIs, including partial least squares (PLS), class modelling analysis, discriminant analysis, multivariate regression and classification [16-22]. However, in many of these studies the number of samples used to build the library was insufficient to capture the natural variability of heparin and no detailed results were reported regarding the detection of contaminants, such as specificity and sensitivity. Monakhova *et al*. [21] proposed models based on linear approaches to detect contamination of non-porcine ruminant material, but with a limited number of samples used for model validations. The same authors in another study, analyzed more than 900 heparin batches, but these data were not used for characterization through explorative analysis with PCA nor or supervised classification purposes [22]. In the present study, several multivariate classification approaches, such as Soft Independent Modelling of Class Analogies (SIMCA), Discriminant Analysis (PCA-DA),

Partial Least Square Discriminant Analysis (PLS-DA) and local classification methods (kNN, BNN and N3), have been systematically tested to detect the presence of samples contaminated by heparin of other animal origin, including ovine heparin and bovine heparin in samples of certified porcine heparin. Classification models were trained and subsequently tested for their predictive capabilities through appropriate validation protocols. A significant number of samples produced by different manufacturers was used to ensure a better representation of the natural biological variability of heparins and thus test the statistical approaches in a more realistic scenario.

2. Material and methods

2.1 Reagents and material

Deuterium oxide 99.9% and 3-(trimethylsilyl)propionic-2,2,3,3-d_4 acid sodium salt (TSP) were purchased from CortecNet and Sigma-Aldrich, respectively. The NMR spectra of the Active Pharmaceutical Ingredient samples had been recorded to build an NMR spectral library of heparin samples, including different sources, different producers and spanned several years of production [from 2011 To 2019]. A total of 76 Porcine Mucosal Heparin (PMH) samples from 19 producers, 56 Bovine Mucosal Heparin (BMH) samples from 4 producers and 17 Ovine Mucosal Heparin (OMH) samples from a single producer were analyzed by 1D-NMR spectroscopy.

In order to train classification models capable of detecting porcine heparin samples contaminated with a low level of contaminant, numerical mixes of ^{1}H NMR spectra were created at 8% BMH and OMH content, according to the following equation:

$$\text{Numerical Mix signals} = \text{PMH signals} \cdot 0.92 + \text{BMH (or OMH) signals} \cdot 0.08$$

The dataset was composed of 76 ^{1}H NMR spectra of pure PMH samples, 69 ^{1}H-NMR spectra comprising the numerical mix of PMH and 8% BMH, and 69 ^{1}H NMR spectra comprising the numerical mix of PMH with 8% OMH. Thirty-eight certified Porcine Heparin samples, 7 Bovine heparin samples extracted from mucosa (BMH) and mixes of porcine heparin contaminated with heparin of the other species at different contamination levels (63 PMH samples containing 5%, <4%, 8%, 10-15% and >16% OMH and 19 PMH samples containing 8%, <4% and >16% BMH) constituted the set of test samples and were therefore used for validation purposes.

2.2 NMR spectroscopy

Solutions for NMR analysis were prepared dissolving 20 mg of sample in 0.6 ml of phosphate buffer solution in deuterium oxide at pH 7.1, the latter was prepared by dissolving 0.36 mmol of sodium dihydrogenphosphate hydrate NaH_2PO_4, 1.14 mmol of disodium hydrogenphosphate dihydrate Na_2HPO_d and 0.03 mmol of deuterate EDTA (d_{16}) in 10 ml of water containing 0.002% of 3-(Trimethylsilyl)propionic-2,2,3,3-

d_4 acid sodium salt (TSP). The pH of the solution was adjusted to 7.1 as necessary. NMR spectra were measured with a Bruker Advance III 600 spectrometer (Bruker, Karlsruhe, Germany) equipped with a 5 mm cryoprobe. Spectra were recorded with a constant presaturation power of 7 Hz at 298 K with standard pulse program (zgcppr) and the following acquisition parameters were used: spectral window of 16 ppm, recycle delay of 12 s, acquisition time of 2 s, pulse length of 90°. Spectra were processed and integrated with Bruker Topspin software version 4.0. After exponential multiplication (line broadening of 0.3 Hz), the spectra were Fourier transformed, phased, baseline corrected and calibrated on the TSP signal. The ^{1}H-^{13}C HSQC spectra were acquired and processed according to the published method [23]. Briefly, ^{1}H-^{13}C HSQC spectra were measured on a Bruker AVANCE III 600 MHz spectrometer equipped with a 5 mm TCI cryoprobe, using the Bruker *hsqcetgpsisp2.2* pulse sequence. The spectra were recorded at 298 K using the following acquisition parameters: number of scans 12, dummy scans 16, relaxation delay 2.5 s, spectral width 8 ppm (F2) and 80 ppm (F1), transmitter offset 4.7 ppm (F2) and 80 ppm (F1), $^{1}J_{C-H}$ = 150 Hz and 1024 points were recorded for each of 240 increments (NUS of 75 % of 320 increments). The FIDs were processed as follows: spectrum size 4096 (F2) and 1024 (F1) (zero-filling in F2 and linear prediction in F1), squared cosine window multiplication in both dimensions and Fourier-transform. The diagnostic heparin building block signals were integrated using Topspin software version 3.5 (Bruker BioSpin, Rheinstetten, Germany) and the heparin composition was computed from the integral values as previously described [23].

2.3 Spectral preprocessing

Spectral intensity binary files and spectra information text files were imported into R using custom scripts [24]. Data cutting was performed according to the chosen spectral region (GAGs signals region: 1.95–2.25, 3.0–3.345, 3.37–3.63, 3.69–4.714, and 4.912–5.75 ppm regions, or part of the anomeric region only, 4.912-5.75 ppm). Spectra were normalized for total area and aligned. The aligned spectra were collected in a bucket table matrix and each column was mean centered and scaled according to autoscaling (standard deviation) or Pareto scaling (square root of the standard deviation). The numerical mixes of the API ^{1}H NMR spectra were created by first adding the spectra of the pure samples to the bucket table and then computing the numerical mix.

2.4 Classification modelling and variable selection

Principal Component Analysis (PCA) and several classification methods were used to analyze and model the NMR spectra [25]. PCA is a well-known unsupervised approach for exploratory data analysis, which projects the data in a reduced hyperspace. This is defined by orthogonal principal components (PCs), which are linear combinations of the original variables (NMR features in this case), with the first principal component having the largest variance, the second principal component having the second-largest variance, and so on. PCA also represents an effective data reduction method, which is used for the subsequent application of supervised classification approaches, such as Discriminant Analysis (DA).

Among traditional classifiers, DA is probably the best-known method, separating samples into classes by minimizing the within-class variance and maximizing the between-class variance [26]. In this study, DA was calculated in the space defined by the first PCs. Another popular discriminant classification approach tested in this study is Partial Least Square Discriminant Analysis (PLS-DA), which is a linear classification method that combines the properties of PLS regression with the discrimination power of a classification technique [27]. In PLS-DA, the relevant sources of data variability are modelled by the Latent Variables (LVs), which again are linear combinations of the original variables. PLS-DA provides quantitative predictions which can be transformed to class labels by means of thresholds [28]. Samples can be assigned to a class when their predictions are higher than a defined limit. When dealing with multiclass tasks, it can happen that PLSDA behaves as a soft classifier and thus samples remain unassigned because they have predictions higher than thresholds for more than one class.

Soft Independent Modelling of Class Analogies (SIMCA) is a class modelling method, also known as one-class classifier [29]. Thus, given a target class, SIMCA is based on a PCA calculated with just the samples belonging to the target class. Then, new samples are predicted in the class according to their distance with respect to the class space defined by the PCA model.

Finally, methods based on local similarity were tested; k-Nearest Neighbours (kNN) uses the concept of analogy for classification and thus a sample is predicted according to the classes of the k closest samples [30]. Similarly to kNN, BNN (Binned Nearest Neighbours) and N3 (N-Nearest Neighbours) classify samples using local information but, with different approaches to identify the local neighbourhood of the sample to be classified [31].

When dealing with supervised classification modelling, validation is a fundamental procedure to ensure the absence of over-fitting and provide a reliable estimation of the ability of the model to predict new samples correctly. In this study, models were validated using the 127 samples included in the test set previously defined, in addition to using cross-validation protocols. These included leave-more-out methods based on 5 cancellation groups. The performance of the classification models was assessed with specific indices and figures of merit. Particularly, sensitivity, specificity and precision of classes were taken into account. These measures define the extent to which each class can be well discriminated in terms of true positives, true negatives and class purity, respectively [32]. Moreover, models were characterised by their Non-Error Rate (NER), also known as balanced accuracy, which is equal to the average of class sensitivities.
Finally, Genetic Algorithms (GA) were used to select specific NMR spectral features which could increase the class discrimination, that is, to identify which spectral intervals are related predominantly to the recognition of contaminated heparin samples. Genetic algorithms mimic the natural selection of a population of so-called chromosomes (models), that reproduces and evolves [33], and starts by randomly defining the chromosomes; a fitness function is then associated to each chromosome (NER in this case). Then, the evolution begins and the best chromosomes '*breed*' with each other, thereby enhancing the selection of the

relevant bits (variables). Different GA strategies exist and, in this study, the frequency of selection of variables (NMR features) was used as a criterion for inclusion of features in the final subset of selected variables [33].

2.5 Software

MATLAB 2017b and 2020b were used to perform the classification models for API samples through several toolboxes, such as the classification toolbox and the PCA toolbox for Matlab [34, 35]. R version 4.0 was used to import and pre-processing of spectral data.

3. Results and discussion

3.1 NMR characterization

The proton spectra of typical PMH, OMH and BMH heparin are shown in **Figure 1**. While the lower level of 6-O-sulfation of BMH affects the profile of its proton spectrum, the spectral profiles of PMH and OMH are very similar, if the weaker acetyl signal of OMH at 2.04 ppm is excluded.

The monosaccharide composition of 76 PMH samples from 19 manufacturers, 56 BMH samples from 4 manufacturers and 17 OMH samples from one manufacturer have been obtained by integration of their HSQC spectra, applying the method described by Mauri *et al.* [23]. The average composition of analyzed batches of each origin is shown in **Table 1** and that of each analyzed batch in **Table S1**-**S3**. The composition of uronic acid and glucosamine residues varies, not only among the three different heparin sources as previously described [10-12], but also within samples of the same origin, in relation to the different purification methods used (**Table 1**). Therefore, the sensitivity of the NMR methods in the detection of non-porcine ruminant material results are strongly affected by the composition of both PMH and contaminants. For example, the main difference between PMH and BMH is the amount of 6-O-sulfation, which equates to 74%-82% and 47%-58% for PMH and BMH, respectively. The sensitivity of the subsequent classification modelling will therefore be affected by the level of heterogeneity within the training samples.

3.2 Pre-processing of data

A preliminary reduction of variables was performed on all API spectra of porcine, ovine and bovine heparin building the Bucket Table and removing all variables related to residual solvent signals or representing just noise. The spectral region selected for the subsequent analysis and classification modelling comprised the combination of the intervals 1.95-2.20 ppm, 3.10-3.34 ppm, 3.72-4.71 ppm and 4.90-5.70 ppm.

The spectra of all samples were aligned as a function of a reference sample, which was chosen at the center of the score plot of the PCA analysis calculated on all samples. Thus, the reference spectrum

represents the closest sample to the data average. Bucketing was applied by considering 4-points means. Pareto data scaling, scaling by the square root of the standard deviation of a mean-centered dataset, was then applied to build the bucket table, which included 4150 variables and was therefore used as the numerical matrix for the calculation of classification models.

The dataset with numerical mixing was thereby composed of 214 samples distributed in 3 classes: 76 pure porcine heparin samples (PMH), 68 samples contaminated with 8% bovine heparin (BMH-mix) and 69 samples contaminated with 8% by ovine heparin (OMH-mix). The number of numerically mixed samples calculated (69 mixes for PMH contaminated by ovine heparin and 69 mixes for PMH contaminated by bovine heparin) was chosen in order to obtain classes that were equally distributed.

3.3 Classification modelling

PCA was initially calculated on pure heparin and numerical mixtures to explore data structure and evaluate the presence of data clustering. The results, however, showed that porcine heparin samples and heparin samples contaminated with heparin from other animal sources at 8% could not be clearly discriminated in the hyperspace defined by the first principal components. Only the third PC (10% of explained variance) could show a marginal discrimination of BMH contaminated samples (**Figure 2**). This result was, nevertheless, to be expected considering the low level of contamination and the similarity of the spectral profiles. Moreover, PCA is suitable for unsupervised data exploration, but it is not aimed at class discrimination. Consequently, supervised classification approaches were applied and compared according to their cross-validation accuracy of detection for samples of porcine heparin contaminated by heparin from another animal sources. In **Table 2** the summary of classification performance in cross-validation is reported.

SIMCA did not provide reliable performance in cross-validation, providing an NER value of 60%. Moreover, many samples (69%) could not be classified when using this class modelling approach because they were placed in the class spaces of more than one class. Classifiers based on local similarity (kNN, N3 and BNN) did not provide satisfactory results either. Both kNN, BNN and N3, however, could better classify BMH-mixed samples, providing consistent sensitivity, specificity and precision for this class in excess of 80%. However, samples of PMH and OMH-mixed classes could not be accurately predicted and therefore the overall classification balanced accuracy of these methods was not suitable for predictive applications, with NER values of 55%, 43% and 64% for kNN, BNN and N3, respectively.

More promising results were obtained with linear discriminant classifiers. Linear Discriminant Analysis (LDA) calculated on the first 13 PCs provided good discrimination among the three classes, with an NER value of 79%. Again, BMH appeared to be the best discriminated class (sensitivity equal to 94% and precision equal to 98%). This result could be expected, since this class is the one which could be best characterized in the previous PCA analysis (**Figure 2**). However, the PMH and OMH-mix classes could also be better modelled than SIMCA and local classifiers. In fact, PCA-LDA provided sensitivity, precision and

specificity values above 68% for both classes. On the other hand, PLS-DA was calculated using 8 latent variables and gave the best classification performance, with an NER value of 92% and very high and balanced sensitivities. In particular, the model correctly classified all samples contaminated with bovine heparin (BMH-mix) and also had high sensitivity for the class of samples contaminated with porcine and ovine heparin, which were often incorrectly classified by the other classification approaches. Owing to its good classification performance, the PLS-DA model was further explored by examining scores (**Figure 3**) and classification coefficients (**Figure 4**). The PLS-DA scores of the first and second latent variables (43% of explained variance) clearly indicated the successful discrimination of the BMH samples, while PMH and OMH samples could be discriminated by higher numbers of latent variables (not shown). Looking at the overall distribution of coefficients, many parts of the NMR spectra seem to be significant for class discrimination; however, for example, all three classes seem to have high (positive or negative) contributions to the discrimination arising from signals associated with high chemical shift (ppm) values (over 5.3). In this case, PLS-DA coefficients suggest that bovine contaminated samples (BMH-mix class) are characterized by the less sulfated sequences (H1 of GlcNS,6OH); the BMH coefficients associated with 5.3-5.4 ppm being negative.

3.3 External validation of PLS-DA model

When comparing the cross-validated results obtained from all the classification approaches considered (**Table 2**), PLS-DA was by far the best method for the discrimination of pure porcine heparin samples from samples contaminated by bovine or ovine heparin. For this reason, PLS-DA was chosen as the method to carry out the subsequent modelling steps. The dataset was refined by excluding three samples which appeared as potential outliers that had been characterised by anomalous NMR spectra in the previous modelling phase. Thus, a refined dataset comprising 211 samples (75 samples of pure porcine heparin, 68 samples of porcine heparin contaminated with bovine heparin and 68 samples of porcine heparin contaminated with ovine heparin) was used to calibrate again a PLS-DA model with 8 latent variables. Test samples were therefore projected in the model and predicted. Test samples which appeared anomalous according to high Hotelling T^2 were discarded (**Figure 5**): heparin samples extracted from BLH and mixtures of heparin samples accounting for more than 80% bovine mucosa heparin appeared to be outliers with respect to the 211 training samples used to calibrate the PLS-DA model, as expected. In addition, samples with a contamination content of more than 20%, therefore higher than that of the training set samples used for the creation of the model, resulted as outliers. These samples were excluded from the test set and the subsequent classification assessment to avoid extrapolation, because they had significantly different NMR features. After this processing, 99 samples were retained in the test set, as described in **Table 3**. When looking at classification results on the test samples, PLS-DA provided satisfactory performances, with NER values of 83% (**Table 4**), thus lower than that calculated in cross-validation (**Table 2**), but still valid. PMH appeared as the best discriminated class (sensitivity, 96%), while the sensitivity of the BMH-mix was lower than that

obtained in cross-validation. Moreover, classification accuracy could be related to the degree of contamination in the test samples. In fact, samples with contamination equal or greater than 5% were all correctly classified by the PLS-DA model (**Table 4**) and only a few samples with contamination equal or lower than 5% were erroneously predicted, such as 7 and 2 OMH-mix samples with contamination lower than 4% and equal to 5%, respectively, and 2 BMH-mix samples with contamination lower than 4%.

Twenty percent of test samples could not be classified by PLS-DA, because they were associated with predicted values lower or higher than the class threshold of more than one class or, lower than the thresholds of all classes. In particular, 12 samples of pure certified heparin and 8 samples with a percentage of contamination less than 5% corresponded to the 20 samples not classified (**Table 3**, **Table 4**). Observing the spectra of these samples (**Figure 6**), it is evident that some are characterized by lower signal intensity in the range 3.9 - 3.8 ppm. This region corresponds to the proton signals 5 and 6 of 6-O-desulfated glucosamine. This means that these heparin samples, even if not indicated as outliers are, nevertheless, characterized by a slightly lower degree of sulfation, compared to the other samples of the test and training sets.

3.4 Variable selection

Variable selection based on GAs was successively applied to identify the most relevant spectral windows for the identification of contamination. This can provide further information on the relationships between the spectral features and the contaminations. Moreover, selection can result in more parsimonious models, that is, classifiers based on fewer variables and therefore less exposed to overfitting, simpler and more easily interpreted.

Genetic algorithm selection was carried out on the dataset composed of pure heparin samples from the three animal sources considered: 76 samples of porcine heparin, 56 samples of bovine heparin and 17 samples of ovine heparin, which represent the same starting dataset from which numerical mixes were calculated for the creation of classification models.

The GAs were calculated testing different pre-processing and alignment methods. Moreover, variables (NMR features) were grouped in 28 contiguous windows (intervals). Variables included in the same window were treated as one input in the selection process. This approach has been demonstrated to improve the selection outcome when dealing with highly correlated data, such as spectra, and avoids overfitting in the selection [33]. The number of windows was chosen by dividing the spectrum into regions in compliance with the number of acquisition points contained in one of the narrowest heparin peaks, specifically the N-acetyl peak at 2.09 ppm. Thus, the full spectrum was divided into windows, each containing consecutive variables.

The GA approach coupled with PLS-DA selected 10 out of the 28 windows, corresponding to 1782 variables. **Figure 7** shows the peaks corresponding to the selected NMR features, in particular some peaks in the anomeric region (5.2, 5.4 and 5.6 ppm) and signals in the range 3.8 - 4.2 ppm. These variables selected

by GA mainly corresponded to those associated with the highest (positive and negative) classification PLS-DA coefficients. The most important coefficients were related to the anomeric region, but also at 3.88 ppm. The magnitude of coefficients is relevant, due to signals originating from H5 + H6 of 6-unsulfated glucosamine, which is more intense for bovine heparin than ovine and porcine samples (**Figure 1**).

After the selection based on pure heparin samples, numeric mixes at 8% were created and, following the same procedure, applied for the mixes calculated for the previous classification models to improve the number of samples. Thus, a dataset consisting of 211 samples (75 PMH, 68 OMH mixes and 68 BMH mixes) was obtained. PLSDA was therefore calculated with this set of data and, considering the 1782 selected NMR features, used to predict the test samples, which were previously predicted with the model calibrated on the full NMR spectra. Classification results achieved on the test samples after variable selection are collected in **Table 6**. The NER, sensitivity and specificity for classes remain approximately the same compared to the previous results (**Table 3** and **Table 4**), but the percentage of samples not classified decreased from 20% to 12%. The test samples not-classified or incorrectly classified were again those characterized by percentages of contamination lower than 8% (**Table 5**).

3.5 Classification of samples with 5% contamination

To assess the sensitivity of the proposed approach to lower contamination levels, the same workflow was applied to a dataset containing numerical mixes with 5% contamination. Starting from the dataset consisting of 149 pure heparin samples (76 porcine, 56 bovine and 17 ovine heparin), a new training set with numerical mix with 5% contamination was generated and used to train another PLS-DA classification model. This new trial was carried out to assess the sensitivity of the proposed approach to contamination levels lower than 8%, which was the one previously tested. The training set comprised 75 samples of porcine heparin, 68 samples of porcine heparin contaminated at 5% by bovine heparin and 68 samples of porcine heparin contaminated at 5% by ovine heparin. The 1782 NMR features previously selected were used as independent variables. The PLS-DA model for the numerical mixes at 5% of contamination performed less well in cross-validation than that obtained by numerical mixes at 8%, with NER values of 80%, while 27% of samples could not be classified. Test samples were again projected in the model and predicted. Outlier test samples were again excluded according to their Hotelling T^2 values and the final test set consisted of 87 samples. Classification results are collected in **Table 6**. The NER value was 86%, with 24% of samples not classified by PLS-DA. This model enables detection of all samples contaminated by heparin from other animal species at 8% of contamination or higher (**Table 6**), while only one test OMH-mix sample with 5% contamination was classified incorrectly.

5. Conclusions

Heparin is the anticoagulant of choice during cardio-pulmonary bypass surgery and for hemodialysis, and, together with its low molecular weight versions, is used globally for the treatment and prevention of thrombosis, requiring almost 100 tons of product per year [36]. Although NMR spectroscopy has proved to be the technique of choice to identify and characterize this complex polydisperse molecule, alone, it cannot fully guarantee the detection of all possible accidental or intentional contamination, including cross-species contamination. The application of chemometric techniques to analyze NMR spectral libraries of heparin enables definition of the characteristic features of each heparin source [16]. Over the past 10 years, numerous articles demonstrating the application of different chemometric classification approaches for the detection of a specific contaminant have been published [17-22]. However, the ability of these models to identify a specific contaminant in a library of fully characterized samples encompassing the real natural variability of the product, has never been fully analysed.

Porcine samples used in this study were characterized by quantitative HSQC spectroscopy and showed that their structural variability related mainly to their different degrees of sulfation and N-acetylation. This variability is induced largely by the different manufacturing processes used to isolate crude heparin from the mucosa and to purify the crude material at the active pharmaceutical ingredient level. Moreover, the mixtures used to build the models (numerical mixtures) were prepared from a series of BMH and OMH batches, which also indicated internal batch-to-batch variability. Among the classification approaches tested, PLS-DA emerged clearly as the most reliable, being able to classify correctly the majority of samples when tested in cross-validation, with more balanced accuracies achieved over the three modelled classes compared to the other classifiers tested. The PLS-DA coefficients could also provide insights into which NMR spectral regions were related most strongly to the class separation and identification. These results were further evaluated by means of variable selection approaches, which provided similar NMR features and thereby enabled the calibration of simpler, but still reliable, PLS-DA classification models. Finally, classification was further validated through an external set of samples with varied levels of contamination and, again, proved able to identify the class of heparin origin. Although the sensitivity of the model is not particularly high, it is, nevertheless, still sufficient to detect adulteration at levels likely to be economically beneficial. These results provide a robust conclusion based on an extensive dataset, confirming earlier findings [21] that were based on fewer samples.

A new technique, time-of-flight secondary ion mass spectrometry, has recently been described as the most sensitive for identifying contamination derived from GAGs of different animal species [37]. Although ToF-SIMS is a promising approach, it has been used to detect PMH contaminated with oversulfated chondroitin sulfate and bovine lung heparin, whose structures differ significantly from that of PMH, thus, present a relatively easy challenge. The advantages of the inherent reproducibility of NMR, the minimal sample manipulation required, minimal potential operator-to-operator variability and the possibility of

automating the entire process for industrial applications, all make NMR a versatile technique highly suited for the quality control environment.

Acknowledgment

This research was supported by “Ronzoni Foundation”, Milan, Italy.

Author contribution

Erika Colombo: Investigation, Methodology, Conceptualization, Writing - original draft. **Lucio Mauri**: Investigation, Methodology, Conceptualization, Writing – review & editing. **Maria Marinozzi**: Investigation, Methodology. **Timothy Rudd**: Conceptualization Writing – review & editing. **Edwin Yates**: Conceptualization, Writing - review & editing. **Davide Ballabio**: Conceptualization, Project administration Supervision, Writing - review & editing. **Marco Guerrini:** Conceptualization, Project administration Supervision, Writing - review & editing.

Table 1 Glucosamine (a) and uronic acid (b) average composition of PMH, OMH and BMH samples. Average, median, standard deviation, minimum and maximum values are indicated. GlcNH2: N-desulfated Glucosamine; GlcNAc: N-acetylated glucosamine; GlcNS,3S; N-sulfated,3-O-sulfated glucosamine; GlcNS: N-sulfated, glucosamine; Glc6S: 6-O-sulfated glucosamine; GlcA: glucuronic acid; GlcA2S: 2-O-sulfated glucuronic acid; IdoA: iduronic acid; IdoA2S: 2-O-sulfated iduronic acid; 2,3-epoxide: 2,3-anydride iduronic acid; GalA: galacturonic acid; x: SO_3^- or H group.

a

		GLUCOSAMINE MONOSACCHARIDES				
		GlcNH2,6x	**GlcNAc,6x**	**GlcNS,3S,6x**	**GlcNS,6X**	**% Glc6S**
PMH	**Average**	1.5	12.9	4.9	81.0	77.5
	δ	0.45	1.24	0.53	1.55	1.82
	Cv	31.03	9.62	10.94	1.91	2.35
	Median	1.5	13.0	4.9	80.8	77.4
	Max	2.8	16.3	6.7	84.2	82.3
	Min	0.6	10.0	3.6	77.6	73.6
OMH	**Average**	0.9	5.8	5.6	88.1	84.7
	δ	0.30	1.03	0.92	1.40	4.64
	Cv	32.20	17.73	16.57	1.59	5.48
	Median	0.9	5.7	5.6	88.3	86.4
	Max	1.3	7.7	7.3	90.0	89.6
	Min	0.5	4.3	3.7	84.8	73.6
BMH	**Average**	0.9	7.1	1.8	90.7	51.7
	δ	0.33	1.41	0.41	1.64	2.30
	Cv	36.06	19.87	23.10	1.81	4.46
	Median	0.8	6.9	1.8	91.0	50.9
	Max	1.8	11.8	2.9	93.1	58.0
	Min	0.4	5.1	1.1	85.5	47.1

b

		URONIC ACID MONOSACCHARIDES					
		GlcA	**GlcA2S**	**IdoA2OH**	**IdoA2S**	**2,3-epoxide**	**GalA**
PMH	**Average**	15.0	nd	8.7	76.1	nd	nd
	δ	1.04	nd	0.87	1.65	nd	nd
	Cv	6.96	nd	10.00	2.16	nd	nd
	Median	14.9	nd	8.9	76.2	nd	nd
	Max	17.5	0.7	10.7	79.6	1.0	3.3
	Min	12.9	0.0	6.9	72.6	0.0	0.0
OMH	**Average**	8.6	nd	5.8	84.9	5.2	0.6
	δ	1.69	nd	0.57	3.97	6.53	0.04
	Cv	19.7	nd	10.0	4.7	126.6	6.61
	Median	7.9	nd	5.7	86.2	5.2	0.6
	Max	12.0	0.0	6.8	87.9	9.8	0.7
	Min	6.8	0.0	4.7	71.4	0.5	0.6
BMH	**Average**	13.5	1.8	4.5	78.8	2.2	3.7
	δ	1.41	0.34	0.65	4.15	1.53	3.37
	Cv	10.46	19.16	14.54	5.26	69.75	92.29
	Median	13.2	1.8	4.4	79.5	1.4	2.2
	Max	17.9	2.7	6.2	84.2	4.8	13.1
	Min	10.3	1.1	3.4	59.6	1.1	0.8

Table 2. Classification measures achieved in cross-validation by classification approaches. Pr: precision; Sn: sensitivity; Sp: Specificity

Model	NER%	Cross-validation parameters			
		Class	**Sn (%)**	**Sp (%)**	**Pr (%)**
SIMCA	60	*PMH*	50	89	45
		BMH-mix	100	95	98
		OMH-mix	30	91	38
kNN	55	*PMH*	38	70	41
		BMH-mix	81	96	90
		OMH-mix	48	66	40
BNN	43	*PMH*	18	58	19
		BMH-mix	87	100	100
		OMH-mix	23	55	20
N3	64	*PMH*	25	87	51
		BMH-mix	96	95	90
		OMH-mix	73	63	48
PCA-LDA	79	*PMH*	74	82	69
		BMH-mix	94	99	98
		OMH-mix	68	86	70
PLS-DA	92	*PMH*	86	95	90
		BMH-mix	100	100	100
		OMH-mix	89	93	85

Table 3. Distribution of test samples in the classes with their contamination level and prediction achieved by PLS-DA.

	Class label	n samples	% of contamination	Samples correctly predicted	Samples not classified
Certificated pure PMH	*PMH*	38	0	25	12
PMH contaminated with OMH	*OMH-mix*	12	< 4%	3	2
		9	5%	4	3
		12	8%	11	1
		17	10-15%	17	0
		2	> 16%	2	0
PMH contaminated with BMH	*BMH-mix*	5	< 4%	1	2
		2	8%	2	0
		2	16%	2	0

Table 4. Classification measures achieved on the test set by PLS.DA. Pr: precision; Sn: sensitivity; Sp: Specificity.

NER		83%	
Not-classified		20%	
Class label	**Sn (%)**	**Sp (%)**	**Pr (%)**
PMH	96	79	69
BMH-mix	71	100	100
OMH-mix	80	97	97

Table 5. Distribution of test samples in the classes with their contamination level and prediction achieved by PLS-DA after GA variable selection.

	Class	n samples	% of contaminant	Samples correctly predicted	Samples not classified
Certificated pure PMH	*PMH*	34		28	5
PMH contaminated with OMH	*OMH-mix*	1	< 4%	3	3
		9	5%	5	5
		12	8	11	0
		15	10-15%	15	0
		5	> 16%	5	0
PMH contaminated with BMH	*BMH-mix*	5	< 4%	0	2
		2	8%	2	0
		2	16%	2	0

Table 6. Distribution of test samples in the classes with their contamination level and prediction achieved by PLS-DA trained on samples with 5% contamination.

	Class label	n samples	% contamination	Samples correctly predict	Samples not classified
Certificated pure PMH	*PMH*	31	0	15	16
PMH contaminated with OMH	*OMH-mix*	12	< 4%	3	2
		9	5%	8	0
		12	8%	12	0
		14	10-15%	14	0
		2	> 16%	2	0
PMH contaminated with BMH	*BMH-mix*	5	< 4%	1	3
		2	8%	2	0

References

[1] E. Gray, J. Hogwood, B. Mulloy, The anticoagulant and antithrombotic mechanisms of heparin, Handb Exp Pharmacol (207) (2012) 43-61.
[2] J.Y. van der Meer, E. Kellenbach, L.J. van den Bos, From Farm to Pharma: An Overview of Industrial Heparin Manufacturing Methods, Molecules 22(6) (2017) 1025.
[3] A.Y. Szajek, E. Chess, K. Johansen, G. Gratzl, E. Gray, D. Keire, R.J. Linhardt, J. Liu, T. Morris, B. Mulloy, M. Nasr, Z. Shriver, P. Torralba, C. Viskov, R. Williams, J. Woodcock, W. Workman, A. Al-Hakim, The US regulatory and pharmacopeia response to the global heparin contamination crisis, Nat Biotechnol 34(6) (2016) 625-30.
[4] 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(6) (2008) 669-75.
[5] J. Fareed, W. Jeske, E. Ramacciotti, Porcine Mucosal Heparin Shortage Crisis! What Are the Options?, Clin Appl Thromb Hemost 25 (2019) 1076029619878786.
[6] F.a.D.A. U.S. Department of Health and Human Services, Heparin for Drug and Medical Device Use: Monitoring Crude Heparin for Quality. https://www.fda.gov/regulatory-information/search-fda-guidance-documents/heparin-drug-and-medical-device-use-monitoring-crude-heparin-quality (accessed 2022-01-08.2022).
[7] C.D. Sommers, D.J. Mans, L.C. Mecker, D.A. Keire, Sensitive detection of oversulfated chondroitin sulfate in heparin sodium or crude heparin with a colorimetric microplate based assay, Anal Chem 83(9) (2011) 3422-30.
[8] D.A. Keire, M.L. Trehy, J.C. Reepmeyer, R.E. Kolinski, W. Ye, J. Dunn, 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(4) (2010) 921-6.
[9] A. Mendes, M.C.Z. Meneghetti, M.V. Palladino, G.Z. Justo, G.L. Sassaki, J. Fareed, M.A. Lima, H.B. Nader, Crude Heparin Preparations Unveil the Presence of Structurally Diverse Oversulfated Contaminants, Molecules 24(16) (2019) 2988.
[10] A. Naggi, C. Gardini, G. Pedrinola, L. Mauri, E. Urso, A. Alekseeva, B. Casu, G. Cassinelli, M. Guerrini, M. Iacomini, V. Baigorria, G. Torri, Structural peculiarity and antithrombin binding region profile of mucosal bovine and porcine heparins, J Pharm Biomed Anal 118 (2016) 52-63.
[11] P.A.J. Mourier, Specific Non-Reducing Ends in Heparins from Different Animal Origins: Building Blocks Analysis Using Reductive Amination Tagging by Sulfanilic Acid, Molecules 25(23) (2020) 5553.

[12] E.A. Yates, F. Santini, M. Guerrini, A. Naggi, G. Torri, B. Casu, 1H and 13C NMR spectral assignments of the major sequences of twelve systematically modified heparin derivatives, Carbohydr Res 294 (1996) 15-27.
[13] K. St Ange, A. Onishi, L. Fu, X. Sun, L. Lin, D. Mori, F. Zhang, J.S. Dordick, J. Fareed, D. Hoppensteadt, W. Jeske, R.J. Linhardt, Analysis of Heparins Derived From Bovine Tissues and Comparison to Porcine Intestinal Heparins, Clin Appl Thromb Hemost 22(6) (2016) 520-7.
[14] L. Mauri, M. Marinozzi, N. Phatak, M. Karfunkle, K. St Ange, M. Guerrini, D.A. Keire, R.J. Linhardt, 1D and 2D-HSQC NMR: Two Methods to Distinguish and Characterize Heparin From Different Animal and Tissue Sources, Front Med (Lausanne) 6 (2019) 142.
[15] L. Mauri, M. Marinozzi, G. Mazzini, R.E. Kolinski, M. Karfunkle, D.A. Keire, M. Guerrini, Combining NMR Spectroscopy and Chemometrics to Monitor Structural Features of Crude Hep-arin, Molecules 22(7) (2017).
[16] T.R. Rudd, D. Gaudesi, M.A. Skidmore, M. Ferro, M. Guerrini, B. Mulloy, G. Torri, E.A. Yates, Construction and use of a library of bona fide heparins employing 1H NMR and multivariate analysis, Analyst 136(7) (2011) 1380-9.
[17] V. Ruiz-Calero, J. Saurina, M.T. Galceran, S. Hernandez-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(4-5) (2002) 259-65.
[18] Q. Zang, D.A. Keire, R.D. Wood, L.F. Buhse, C.M. Moore, M. Nasr, A. Al-Hakim, M.L. Trehy, W.J. Welsh, Class modeling analysis of heparin 1H NMR spectral data using the soft independent modeling of class analogy and unequal class modeling techniques, Anal Chem 83(3) (2011) 1030-9.
[19] Q. Zang, D.A. Keire, R.D. Wood, L.F. Buhse, C.M. Moore, M. Nasr, A. Al-Hakim, M.L. Trehy, W.J. Welsh, Determination of galactosamine impurities in heparin samples by multivariate regression analysis of their (1)H NMR spectra, Anal Bioanal Chem 399(2) (2011) 635-49.
[20] Y.B. Monakhova, B.W. Diehl, Combining 1H NMR spectroscopy and multivariate regression techniques to quantitatively determine falsification of porcine heparin with bovine species, J Pharm Biomed Anal 115 (2015) 543-51.
[21] Y.B. Monakhova, B.W.K. Diehl, J. Fareed, Authentication of animal origin of heparin and low molecular weight heparin including ovine, porcine and bovine species using 1D NMR spectroscopy and chemometric tools, J Pharm Biomed Anal 149 (2018) 114-119.
[22] Y.B. Monakhova, B.W.K. Diehl, Retrospective multivariate analysis of pharmaceutical preparations using (1)H nuclear magnetic resonance (NMR) spectroscopy: Example of 990 heparin samples, J Pharm Biomed Anal 173 (2019) 18-23.
[23] L. Mauri, G. Boccardi, G. Torri, M. Karfunkle, E. Macchi, L. Muzi, D. Keire, M. Guerrini, Qualification of HSQC methods for quantitative composition of heparin and low molecular weight heparins, J Pharm Biomed Anal 136 (2017) 92-105.
[24] R Core Team, R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing, 2022.
[25] R. Bro, A.K. Smilde, Principal component analysis, Anal. Methods 6(9) (2014) 2812-2831.
[26] G.J. McLachlan, Discriminant Analysis and Statistical Pattern Recognition, John Wiley & Sons, Hoboken, NJ, USA, 1992.
[27] M. Barker, W. Rayens, Partial least squares for discrimination, Journal of Chemometrics 17(3) (2003) 166-173.
[28] N.F. Pérez, J. Ferré, R. Boqué, Calculation of the reliability of classification in discriminant partial least-squares binary classification, Chemometrics and Intelligent Laboratory Systems 95(2) (2009) 122-128.
[29] R.G. Brereton, One-class classifiers, Journal of Chemometrics 25(5) (2011) 225-246.
[30] T. Cover, P. Hart, Nearest neighbor pattern classification, IEEE Transactions on Information Theory 13(1) (1967) 21-27.
[31] R. Todeschini, D. Ballabio, M. Cassotti, V. Consonni, N3 and BNN: Two New Similarity Based Classification Methods in Comparison with Other Classifiers, J Chem Inf Model 55(11) (2015) 2365-74.
[32] D. Ballabio, F. Grisoni, R. Todeschini, Multivariate comparison of classification performance measures, Chemometrics and Intelligent Laboratory Systems 174 (2018) 33-44.
[33] R. Leardi, A. Lupiáñez González, Genetic algorithms applied to feature selection in PLS regression: how and when to use them, Chemometrics and Intelligent Laboratory Systems 41(2) (1998) 195-207.

[34] D. Ballabio, A MATLAB toolbox for Principal Component Analysis and unsupervised exploration of data structure, Chemometrics and Intelligent Laboratory Systems 149 (2015) 1-9.
[35] D. Ballabio, V. Consonni, Classification tools in chemistry. Part 1: linear models. PLS-DA, Analytical Methods 5(16) (2013) 3790-3798.
[36] S.L. Taylor, J. Hogwood, W. Guo, E.A. Yates, J.E. Turnbull, By-Products of Heparin Production Provide a Diverse Source of Heparin-like and Heparan Sulfate Glycosaminoglycans, Sci Rep 9(1) (2019) 2679.
[37] A.L. Hook, J. Hogwood, E. Gray, B. Mulloy, C.L.R. Merry, High sensitivity analysis of nanogram quantities of glycosaminoglycans using ToF-SIMS, Communications Chemistry 4(1) (2021) 67.