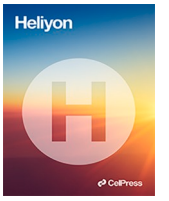
Heliyon 9 (2023) e16331

Contents lists available at ScienceDirect

Heliyon

journal homepage: www.cell.com/heliyon



Research article

A machine learning approach for identifying variables associated with risk of developing neutralizing antidrug antibodies to factor VIII

Atul Rawal [a], Christopher Kidchob [a], Jiayi Ou [a], Osman N. Yogurtcu [b], Hong Yang [b], Zuben E. Sauna [a,*]

[a] *Hemostasis Branch, Division of Plasma Protein Therapeutics, Center for Biologics Evaluation and Research, Food and Drug Administration, USA*
[b] *Division of Analytics and Benefit Risk Assessment, Office of Biostatistics and Pharmacovigilance, Center for Biologics Evaluation and Research, Food and Drug Administration, USA*

A R T I C L E I N F O

Keywords:
Factor VIII
Hemophilia
Immunogenicity
Machine learning
Explainable AI (XAI)
Anti-drug antibodies

A B S T R A C T

A key unmet need in the management of hemophilia A (HA) is the lack of clinically validated markers that are associated with the development of neutralizing antibodies to Factor VIII (FVIII) (commonly referred to as inhibitors). This study aimed to identify relevant biomarkers for FVIII inhibition using Machine Learning (ML) and Explainable AI (XAI) using the My Life Our Future (MLOF) research repository. The dataset includes biologically relevant variables such as age, race, sex, ethnicity, and the variants in the *F8* gene. In addition, we previously carried out Human Leukocyte Antigen Class II (HLA-II) typing on samples obtained from the MLOF repository. Using this information, we derived other patient-specific biologically and genetically important variables. These included identifying the number of foreign FVIII derived peptides, based on the alignment of the endogenous FVIII and infused drug sequences, and the foreign-peptide HLA-II molecule binding affinity calculated using NetMHCIIpan. The data were processed and trained with multiple ML classification models to identify the top performing models. The top performing model was then chosen to apply XAI via SHAP, (SHapley Additive exPlanations) to identify the variables critical for the prediction of FVIII inhibitor development in a hemophilia A patient. Using XAI we provide a robust and ranked identification of variables that could be predictive for developing inhibitors to FVIII drugs in hemophilia A patients. These variables could be validated as biomarkers and used in making clinical decisions and during drug development. The top five variables for predicting inhibitor development based on SHAP values are: (i) the baseline activity of the FVIII protein, (ii) mean affinity of all foreign peptides for HLA DRB 3, 4, & 5 alleles, (iii) mean affinity of all foreign peptides for HLA DRB1 alleles), (iv) the minimum affinity among all foreign peptides for HLA DRB1 alleles, and (v) *F8* mutation type.

1. Introduction

In the last two decades therapeutic proteins have proved to be very successful as they address serious clinical conditions, are targeted, and have limited side effects. However, immune responses to the drug (immunogenicity) are a unique feature of this class of

* Corresponding author.
E-mail address: zuben.sauna@fda.hhs.gov (Z.E. Sauna).

https://doi.org/10.1016/j.heliyon.2023.e16331
Received 18 October 2022; Received in revised form 21 April 2023; Accepted 12 May 2023
Available online 23 May 2023

medications. The unwanted immune response to a therapeutic protein affects both the efficacy and safety of the product.

Replacement proteins have revolutionized the management of individuals with many genetic diseases. This is clearly the case with hemophilia A where Factor VIII (FVIII) replacement proteins have significantly decreased mortality and improved the quality of life. The modern management of hemophilia involves prophylaxis, i.e., the routine infusions of FVIII drug products. These products are either purified from human plasma or manufactured using recombinant DNA technology [1]. Unfortunately, as with most therapeutic proteins, hemophilia A patients elicit neutralizing anti-drug antibodies (referred to as inhibitors in the hemophilia literature). Inhibitors, which occur in ~30% of patients treated with FVIII, complicate the management of the disease, increase the cost of medical care and is a burden on patients and their caregivers. Hemophilia A is a genetic deficiency in clotting factor VIII, which almost always afflicts males as it is an X-linked recessive trait. Females can be carriers of *F8* gene defects but are rarely treated. However, recent studies of female hemophilia A carriers status show that many have bleeding complications that remain underrecognized and untreated [2]. As treatment of female carriers of hemophilia becomes more routine, it will be interesting to study the immunogenicity consequences in these patients.

Inhibitors to FVIII replacement therapy have been extensively studied. Numerous product and patient-related variables have been associated with the risk of inhibitor formation. The patient-related variables include the mutation(s) [3] and polymorphisms in the *F8* gene [4–6], the patient's major histocompatibility complex (MHC), also known as human leukocyte antigens (HLA) [7], race [8], ethnicity [9], and age at first treatment [10]. Product related variables include the source of the FVIII therapeutic (plasma derived or recombinant) [11] and whether first- second- or third-generation recombinant products were used [12]. These risk-factors for inhibitor development were identified in individual studies which often had small sample sizes.

The large dataset generated by the My Life Our Future (MLOF) project provides a valuable resource for applying Artificial Intelligence (AI) based tools for identifying variables. An AI-based approach has several advantages. AI-based tools can uncover patterns, generate robust models, and make data-based prediction by learning from the given training data. These models also drastically reduce the computational time and cost associated with traditional regression/classification approaches. While statistical approaches inherently deduce relations from data, AI approaches are aimed at making robust accurate predictions based on the data. The key difference between the two approaches also lies in their purpose, statistical approaches are programmed with rules to infer relations from the data, whereas AI approaches learn from the data-itself.

For prediction of inhibitor development, classification models are suitable as they can make binary classification predictions based on the datasets [13]. The top ML classification models include Random Forest (RF), Logistic Regression (LR), Gradient Boosting (GB), Light Gradient Boosting (LGBM), Extreme Gradient Boosting (XGB), and CatBoost (CB). In addition to ML classifiers, XAI based models provide the added feature of listing and ranking the input variables in accordance with their importance to the model. That is, they list the features based on how much of an impact each variable makes on the prediction individually.

In this study we used XAI to identify and rank variables associated with inhibitor development in HA. The following were identified as the top five variables (potential biomarkers) associated with risk of inhibitor development in individual hemophilia A patients: baseline activity of the FVIII protein, mean affinity of all foreign peptides for HLA DRB 3, 4, & 5 alleles, mean affinity of all foreign peptides for HLA DRB1 alleles, the minimum affinity among all foreign peptides for HLA DRB1 alleles, and *F8* mutation type. These variables are consistent with those previously associated with inhibitor development (see above). The use of XAI, however, provides a robust and ranked identification of variables that could be predictive for inhibitors to *F8* which is used to treat hemophilia A patients.

2. Materials & methods

2.1. Data sources

The MLOF program is the result of a collaboration between the ATHN, NHF, and Bloodworks Northwest, with support of Bioverativ through June 2018. Subjects and/or their parents gave written informed consent for inclusion of their samples and data in the MLOF Research Repository. Phenotypic data on MLOF Research Repository subjects was abstracted from the ATHNdataset collected from participating hemophilia treatment centers around the United States, including demographic, phenotypic, and genomic data.

We have previously HLA typed 1,000 samples from the MLOF repository [7] and we used this information in our analyses.

2.2. Methods

We used an XAI approach to identify variables (potential biomarkers) associated with inhibitors in hemophilia A patients. In addition to variables present in the MLOF dataset; race, age, sex, disease severity, baseline activity of the FVIII protein, and mutation, we computed biologically important variables such as the number of identified foreign peptides derived from the FVIII drug in individual patients and the foreign-peptide HLA-II affinity. The baseline activity was provided to the authors in the MLOF dataset. In the MLOF program, the hemophilia treatment centers locally collect this information from the volunteers. The baseline activity test is performed prior to treatment with factor replacement products. As our dataset includes information from hemophilia patients across many generations, the date of this test varies significantly, with median year of test being 2007. Consequently, the performed activity test could be of several different methods, including one-stage clotting assays, chromogenic assays, and immunoassays. The baseline activity of the FVIII protein is provided in units of percentage. The percentage represents the amount of Factor VIII activity in a patient's plasma compared to a reference sample. The reference sample is usually a normal plasma sample that has been standardized to contain a known amount of Factor VIII activity. In MLOF, less than 1% of normal level of factor VIII is considered as having severe disorder, between 1% and 5% is considered as having moderate disorder, and between 5% and 50% as having mild disorder. The data

were processed and trained with multiple ML classification models to identify top performing models. We included the following ML classification models: Logistic Regression (LR), Random Forest (RF), Gradient Boosting (GB), Light Gradient Boosting (LGBM), Extreme Gradient Boosting (XGB), CatBoost (CB), and Convolutional Neural Network (CNN) classifiers. The dataset was randomly divided at 80%–20% split for training and validation for the models. Upon completion of the training and validation the models were evaluated for performance using metrices for accuracy, precision, recall, and F1-score. The top performing model was then chosen to apply XAI via SHAP, (SHapley Additive exPlanations) to identify the variables critical for the prediction of Factor VIII inhibitors in a hemophilia A patient.

The MLOF dataset was preprocessed and to have only the biologically relevant variables used for the ML classification models as shown in the schematic in Fig. 1(A). The final dataset was filtered down from 7151 patients to include information on 940 patients as there were numerous, irrelevant variables and variables with unknown values, thus patients with missing information were excluded. 11 patients from the 940 patients did not have race information, but these were included in the study as all the other variables were available for these patients. Within these 940 samples we derived other patient-specific variables such as the number of foreign peptides, based on the alignment of the endogenous FVIII and infused FVIII drug sequences, and the foreign-peptide HLA-II affinity (Fig. 1(B)). To obtain the list of foreign peptides that can be formed from the drug sequence and patient FVIII sequence, we utilized R's MSA library to preform multisequence alignment. Patients with unknown drug treatments were removed from this affinity analysis. For the patients, variants in the *F8* gene may or may not have functional consequences. The HLA-typed dataset included information for DRBs 1,3,4, and 5, thus the foreign peptide affinity for these alleles were calculated and used as variables. Moreover, previous experimental studies have shown that DRB1 alleles are the ones most relevant for FVIII immunogenicity. For instance, a major Histocompatibility Complex (MHC) Associated Peptide Proteomics study showed that most FVIII derived peptides bind to HLA DRB1 proteins [14]. We classified variants as functional if they are amino acid substitutions that do not affect the activity of the FVIII protein (such substitutions were previously also referred to functional variants). These substitutions refer to both missense and nonsense variants and does not include synonymous substitution events. Non-functional modifications are changes to the primary sequence of FVIII negatively affect protein function. From an immunological perspective, both functional and non-functional variants represent a mismatch between the sequences of the endogenous and infused FVIII and can potentially elicit an immune response. The MLOF dataset was generated using materials from hemophilia A patients who volunteered to participate in this program. Thus, the data set only includes individuals who carry the pathogenic variant. Control individuals who carry benign F8 variants would not manifest the disease, would not be treated with replacement FVIII and thus immune responses to the treatment would be a moot point. That being said, it is plausible that hemophilia A patients could carry a benign F8 variant in addition to a pathogenic variant. We therefore identified those mutations that are functional using the database of Human Genome Variations [https://hgv.figshare.com]. Functional and non-functional variants are presented as different sets of variables in the model. Thus, in our schema, differences in the wild type and the patient sequence are used to flag the sequence in the infused FVIII drug as foreign. Using key terms to determine the type of modification, we wrote several functions to apply these changes to the patient sequence. The modified DNA sequences are then retranslated into RNA using a hash table, ending the sequence once a stop codon is translated or the sequence has reached the final residue position. After aligning the mutated sequences of the patient and their listed drug, we index the residues of the drug sequence

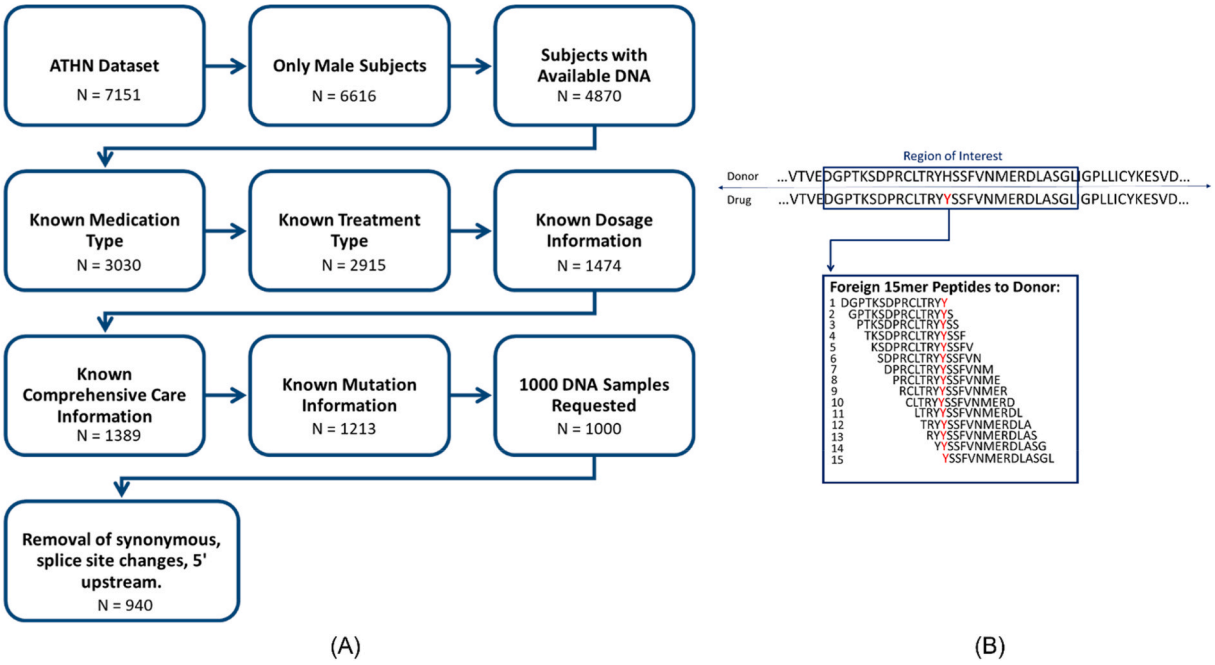


Fig. 1. Schematics for (A) data filtering process and (B) the algorithm for finding the list of foreign 15mer peptides derived from the infused FVIII.

that differ from the patient sequence. Using these indexed positions, we perform a sweep along the drug sequence to generate 15mers, making selections of 15 connected residues within the drug sequence where there is at least one different residue in the drug sequence compared to the patient sequence. From there we store a list of the generated 15mers for each patient and generate a count of all 15mers with a mismatch between the infused (drug) and endogenous FVIII sequences.

When combined with HLA data for the patient, we predict the binding affinity for each foreign peptide and the patient's HLA-II molecules using NetMHCIIpan 3.2. For each patient for whom we had generated HLA data, we parsed DRB1, DRB3, DRB4, and DRB5 alleles information. Participant data for DRB3, DRB4 and DRB5 were limited, as not all participants had the data for all three DRBs. Therefore, we grouped the data for the three DRBs as a single variable to ensure that data was available for each participant. We used the metric of percentile rank instead of the binding affinity value. The specific numeric values for nanomolar that predict strong binding vary between HLA alleles such that a value that may be a strong binder for one allele may be considered a weak binder for another. This makes it difficult to do comparisons between the alleles, thus, to standardize across all alleles we used percentile ranks of the nanomolar binding affinity for each allele. By using this method of standardization, we can designate the top 10% rank and higher as a threshold for strong binders and have a consistent metric throughout. To optimize calculation time, we pre-calculate by generating a list of unique foreign peptides across all the patients with HLA data for each DRB alleles. Once the affinity values for each HLA patient have been calculated, we find the minimum (highest binding/top percentile rank), median, and average percentile ranks for both their DRB1 alleles and the combination of their DRB3, DRB4, and DRB5 alleles.

Once the dataset was filtered, we imported it into our python notebook using pandas DataFrame. We performed data visualization using different plots imported from the matplotlib. pyplot libraries. Since the dataset combines both numerical and categorical variables, the entire dataset was numerically encoded to datatype float64. The new encoded dataset was then randomly split for testing and validation using an 80%–20% split with 80% of the dataset used for training the model and 20% used for validation. The training dataset was then applied to the different ML classification models mentioned earlier to compare and pick the top performing model.

We generated ML classification models by importing libraries such as Scikit-learn (sklearn), pandas, NumPy, and Keras into a Python notebook. LR, RF, and GB classifiers were imported from sklearn. LGBM, XGB and CB classifiers were imported from LightGBM, XGBoost and CatBoost respectively, and the CNN classifier was imported from Keras. Other libraries such as auto encoders and SHAP were installed directly for the notebook using pip. We used metrics of accuracy, precision, recall, and F1-score for evaluating the performance of the models. The definitions for these metrices are given below:

- ***Accuracy*** **-** The base metric used for model evaluation is often *Accuracy*, describing the number of correct predictions over all predictions.

$$\text{Accuracy} = \frac{True\ Positive + True\ Negative}{True\ Positive + True\ Negative + False\ Positive + False\ Negative}$$

- ***Precision*** **–** Precision is a measure of how many of the positive predictions made are correct.

$$\text{Precision} = \frac{True\ Positive}{True\ Positive + False\ Positive}$$

- ***Recall*** **-** Recall is a measure of how many of the positive cases the classifier correctly predicted, over all the positive cases in the data.

$$\text{Recall} = \frac{True\ Positive}{True\ Positive + False\ Negative}$$

- ***F1-Score*** **-** F1-Score is a measure combining both precision and recall. It is generally described as the harmonic mean of the two. It is between 0 and 1, where 0 is the worst score and 1 indicates that the model predicts each observation correctly.

$$\text{F1 Score} = 2 * \frac{Precision * Recall}{Precision + Recall}$$

Once the models were evaluated, the top performing model was chosen for further analysis using XAI. XAI was applied to the top performing model using SHAP, which was installed and imported using pip. SHAP is an open-source post-hoc explanation method used for explaining ML and DL models. It provides explanations via feature relevance where an importance value is assigned to each feature for specific predictions. SHAP also provides a ranking of the input variables in the order of their impact on the model. A linear explainer was used for explaining the logistic regression model; a Kernel explainer was used for the CNN model; and the tree explainer was used for explaining the rest of the models. Explainability of the models were visualized using a feature relevance plot, and a beeswarm plot.

3. Results

3.1. Characteristics of HLA typed patients

The MLOF dataset provided us with genetic and clinical information for 7,151 hemophilia A patients. DNA samples from 1,000 of

these patients were provided to us for HLA typing. The HLA typing has been reported previously [7]. The characteristics of the patient population included in the analysis are depicted in Table 1. A prior study by McGill e. t al., used the same dataset used in the current study to investigate the difference between the mild, moderate, and severe cases of the disease to demonstrate how the patient populations compares with those described in other studies of inhibitor risks. The variables used in the study are shown in Table 2. In addition to the variables provided by the MLOF program we computed (see Methods for details) the following variables for each of the patients: the number of foreign peptides, and the peptide HLA-II binding affinity (expressed as a percentile rank) for each of the foreign peptides.

3.2. Variables included in the analysis and inhibitor status

To provide an overview of our dataset with respect to inhibitor status we plotted the fraction of the population positive for inhibitors for each of the variables. These variables have been ranked based on the fraction of individuals with that variable who were inhibitor positive (Fig. 2). This representation is a visualization tool that is not based on any statistical tests. Variables associated with a high fraction (>30%) of inhibitor positive subjects include race (black – 36.84%), severity (severe – 35.50%), number of foreign peptides (30–500–33.01%) and (1000–2336–43.53%), and mutations (Intron 1–43.75%, Intron 22–40.07%, nonsense – 42.67%, and large structural change – 42.72%).

3.3. Numerical and categorical pair plots

Pair plots are an easy and reliable method for data analysis that allow us to visualize the distribution of each individual variable and the relationships between each variable within the dataset. These plots produce a relationship matrix for all the variables within the dataset. For the MLOF dataset we generated two sets of pair plots, one for the categorical variables, and the other for the numerical variables. Since pair plots only work with numerical variables, we encoded the categorical features as Pandas datatype float64, which is the "floating point number" data construct used for data storage and manipulation within the python programming language, and plotted them. From the generated pair plots no direct correlations were inferred between any pair of the variables (Supplementary Figure 1 and Fig. 2).

Table 1
Participant characteristics for the dataset. (11 patients did not have race information available.)

Variable	Inhibitor Negative		Inhibitor Positive		Total	
	Number	Percentage	Number	Percentage	Number	Percentage
Number of participants	696	74.04%	244	25.96%	940	100%
Race						
White	573	76.10%	180	23.90%	753	80.11%
Black	72	63.16%	42	36.84%	114	12.13%
Asian	28	66.66%	14	33.33%	42	4.46%
Others	16	80%	4	20%	20	2.12%
Disease Severity						
Mild	204	91.89%	18	8.11%	222	23.62%
Moderate	125	83.89%	24	16.11%	149	15.85%
Severe	367	64.50%	202	35.50%	569	60.53%
Mutation on F8						
Intron 1	9	56.25%	7	43.75%	16	1.70%
Intron 22	160	59.93%	107	40.07%	267	28.40%
Substitution	436	84.99%	77	15.01%	513	54.57%
Frameshift	393	70.75%	31	29.25%	106	11.28%
Missense	43	89.73%	45	10.27%	438	46.60%
Nonsense	4	57.33%	32	42.67%	75	7.98%
Small structural change	181	80.00%	1	20.00%	5	0.53%
Large structural change	75	57.28%	135	42.72%	316	33.62%
Biological age group						
0–2 years	0	0	4	100%	4	0.42%
3–12 years	120	61.22%	76	38.77%	196	20.85%
13–18 years	141	78.77%	38	21.23%	179	19.04%
19–29 years	169	71.61%	67	28.39%	236	25.10%
30–49 years	163	81.09%	38	18.91%	201	21.38%
50–74 years	93	82.30%	20	17.70%	113	12.02%
75 years	10	90.91%	1	9.09%	11	1.17%
Number of foreign peptides group						
FP 15	358	78.00%	101	22.00%	459	48.83%
FP (15–30)	104	82.54%	22	17.46%	126	13.40%
FP (30–500)	140	66.99%	69	33.01%	209	22.23%
FP (500–1000)	46	75.41%	15	24.59%	61	6.49%
FP (1000–2336)	48	56.47%	37	43.53%	85	9.04%

Table 2
Variables used for the machine learning (ML) classification models.

Variables
Race
Mutation
Biological Age Group
Number of Foreign Peptides
Baseline Activity for *F8* Protein
Foreign Peptide Minimum MHC-II Affinity – DRB1
Foreign Peptide Mean MHC-II Affinity – DRB1
Foreign Peptide Minimum MHC-II Affinity – DRB345
Foreign Peptide Mean MHC-II Affinity – DRB345

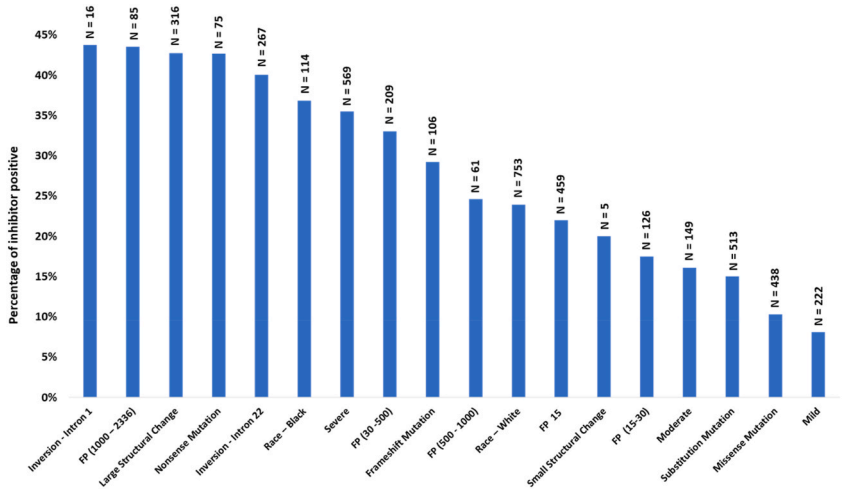


Fig. 2. Fractions of inhibitor positive individuals grouped by variable of interest.

3.4. Evaluating machine learning tools for use in identifying variables associated with inhibitors

We evaluated the following ML tools to select the one best suited for the analyses: RF, LR, LGB Method, EGB, GB, CB, and CNN classifiers. These AI tools were used for our dataset as they are the top performing ML classification models, and we are working on a binary classification problem of whether a patient would develop inhibitors or not [13]. These tools were assessed for accuracy, precision, recall and the F1-Score (see Methods for a detailed description of each of these parameters). As mentioned in the Methods section, the dataset was split into an 80–20 ratio for training and validation. Once the model was trained, it was validated with 20% of the samples using the performance markers mentioned previously. These evaluation metrices were calculated for all the models using the pipeline tool within the sklearn library in python. The results of the model comparison are depicted in Table 3 along with the confusion matrix for all six ML classifiers (Fig. 3). Confusion matrices display the number of True negatives (TN), True Positives (TP), False Negatives (FN) and False Positives (FP) on the diagonals. The left diagonal (top left and bottom right) stands for the number of true negatives (TN) on the top left and the true positives (TP) on the bottom right, while the right diagonal (top right and bottom left) represents the number of false negatives (FN) on the top right and the false positives (FP) on the bottom left. All ML tools performed well (AUC>0.7 for accuracy, precision, recall and F1-Score), however RF and LGBM performed best across all criteria. We used the LGBM for detailed analyses of the dataset since SHAP is readily available for LGBM with a Tree Explainer.

Table 3
Model comparison with calculated statistical performance for difference machine learning and deep learning classifiers.

Model	Cross Validation Accuracy	Precision	Recall	F1-Score	Total
Random Forest	0.7420	0.99	0.99	0.99	3.712
Light Gradient Boosting	0.7354	0.99	0.99	0.99	3.705
Cat Boosting	0.7420	0.89	0.89	0.88	3.402
Gradient Boosting	0.7274	0.85	0.84	0.82	3.237
Extreme Gradient Boosting	0.7314	0.81	0.81	0.80	3.151
Logistic Regression	0.7513	0.68	0.72	0.63	2.781
Convolutional Neural Network	0.7340	0.33	0.02	0.03	1.114

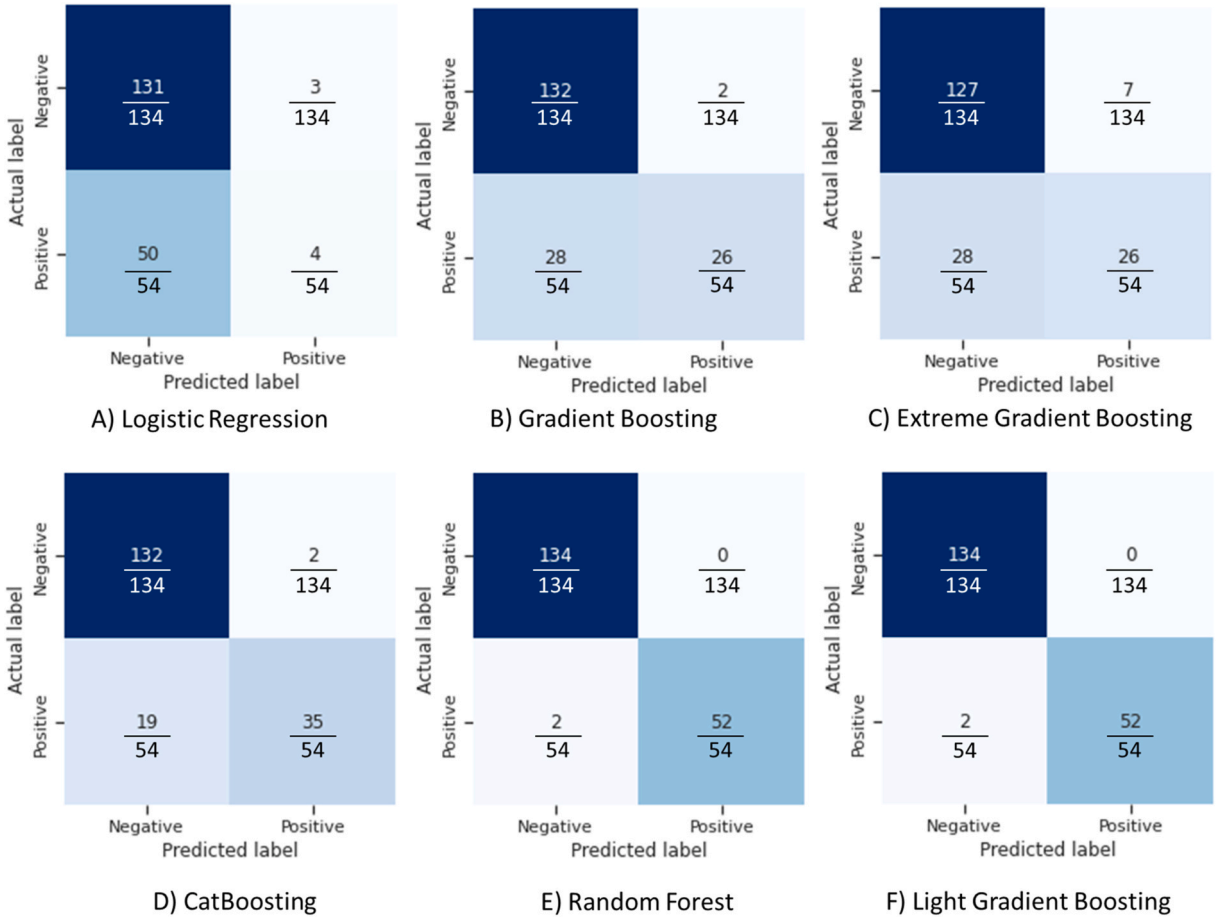


Fig. 3. Confusion matrix for all six ML classifiers.

3.5. *Using an explainable Artificial Intelligence (XAI) based tool for identifying variables associated with inhibitors to FVIII*

We used SHAP to identify and rank the variables associated with inhibitors to FVIII (Fig. 4(A)). SHAP allows for the calculation of SHAP values which represent the impact each variable has on the prediction made by the model.

The SHAP value plot shows the relationship between the predictor features (variables) and the target classification (FVIII inhibitors determined in the Bethesda assay). For the subjects in the MLOF cohort, inhibitors were estimated at the individual Hemophilia Treatment Centers. In most instances a Bethesda Assay with the Nijmegen modification is used [15,16]. The assay involves mixing

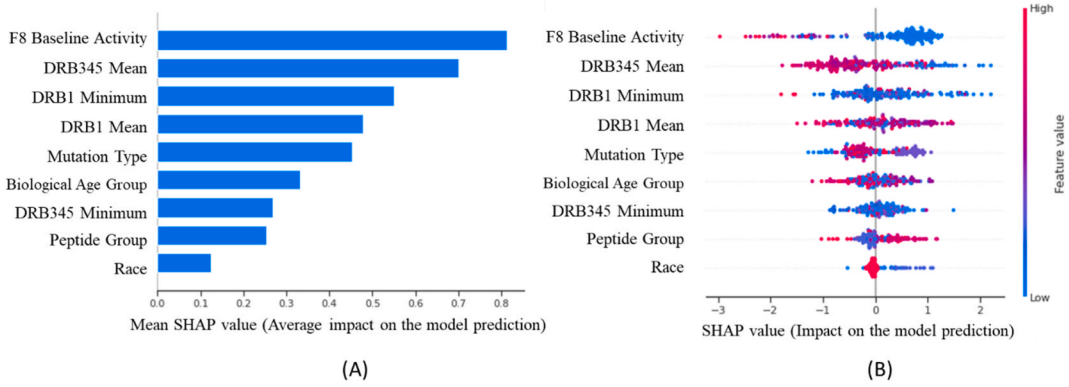


Fig. 4. SHAP plots generated for LGBM classifier which had a cross-validation accuracy of 0.7354 and recall, precision and F1 scores of 0.99. (A) feature relevance, and (B) beeswarm feature relevance, which highlights the ranked variables from top to bottom.

equal volumes of patient plasma and normal pooled plasma and FVIII activity is compared to that of a control consisting of normal pooled plasma with an equal volume of buffer. A Bethesda Unit is defined as the amount of inhibitor that will inactivate half of the factor present in an equal mixture of patient and normal pooled plasma following incubation at 37 °C for 2 h. The Nijmegen modification of the assay includes the use of FVIII-deficient plasma or 4% BSA in lieu of buffer in the control sample and incorporation of buffered normal pooled plasma. The International Society on Thrombosis and Haemostasias (ISTH) recommends the use of the Nijmegen modification of the Bethesda Assay [17]. The assay is now often performed using the Nijmegen modification of the original method to improve test accuracy. In our dataset, close to 90% of all hemophilia patients have inhibitor tests performed after 2000. Considering that the Nijmegen modification for FVIII was introduced in early 1990, we expect the assay method for the large majority of these tests in our dataset is Nijmegen. The SHAP value plot depicts the effect in for positive and negative directions, i.e., it shows how much of an impact, both positive and negative, the specific variable makes on the target classification. The plot is made up of all the sample points from the training dataset. We ranked all the variables from top to bottom according to the impact they have on the inhibition classification. The mean absolute feature (variable) values are shown as ranked vertical bar to the right, which shows the color gradient for the values going from blue for low to red for high, which corresponds to the low and high values for the individual variables as they have been numerically encoded for the model. For example, the values for variable F8_activity_baseline ranges from 0%–47%. The higher valued features (red) have a lesser impact (negative x-axis value below 0) on the inhibition, and the lower valued features (blue) have a larger positive impact (positive x-axis), as seen on Fig. 4(B). This corresponds to a higher inhibition rate for patients with smaller FVIII baseline activity and a severe disease.

We identified the following as the top five variables (potential biomarkers) associated with inhibitor development in individual hemophilia A patients: baseline activity of the FVIII protein, mean affinity of all foreign peptides for HLA DRB 3, 4, & 5 alleles, minimum affinity among all foreign peptides for HLA DRB1 alleles, mean affinity of all foreign peptides for HLA DRB1 alleles, and F8 mutation type.

4. Discussion

It has been recognized for over three decades that inhibitors hamper the effective management of hemophilia A patients who receive FVIII replacement therapy [18]. Numerous investigations have sought to identify risk-factors associated with FVIII inhibitors. These studies have yielded several patient and product-related risk-factors (see introduction). Unfortunately, these studies are often fragmentary (studying a single or a few risk factors), use diverse methodologies to measure the variables and many of the clinical studies lack power. To address some of these drawbacks large prospective studies were designed and executed to evaluate specific hypotheses, such as the relative immunogenicity risk of different FVIII products [11,12].

An evaluation of 574 previously untreated children with severe hemophilia A found that plasma-derived products conferred a risk of inhibitor development that was similar to the risk with recombinant products. However, second-generation full-length recombinant FVIII products were associated with an increased risk, as compared with third-generation FVIII products [12]. A randomized trial assessed the incidence of FVIII inhibitors in 264 previously untreated children treated with either plasma derived or recombinant FVIII products [11]. This study concluded that treatment with recombinant FVIII resulted in a higher incidence of inhibitors.

The two carefully designed randomized trials described above are the exception rather than the rule. Moreover, these hypotheses driven studies generally focus on a single variable (e.g., product type or *F8* mutation). Numerous additional variables which have been reported to be associated with the clinical risk of inhibitor development have never been evaluated in large, controlled studies. For instance, the putative association of genetic defects in the *F8* gene and inhibitor status has come from many small studies and a meta-analysis that included 30 independent studies and 5,383 patients [3].

While the role of *F8* gene defects and functional variants in the pharmacogenetics of inhibitor development in hemophilia A have received the most attention, other genetic factors could also play a role. For instance, several studies suggest that the HLA repertoire of the patient may play an important role in the immune response to FVIII products and other therapeutic proteins [7,19,20]. The simultaneous, unbiased evaluation and rank ordering of the various variables associated with eliciting inhibitors to FVIII products remains an important unmet need. In this study we have used an AI-based approach to interrogate a large dataset, of clinical and genetic variables associated with hemophilia A patients which was provided by the MLOF consortium.

MLOF provided information on genetic, clinical, and demographic variables and many of these have been associated with inhibitor status in hemophilia A patients. These variables include variants in the *F8* gene [4–6], FVIII product used [11,12], race [8], and age [10]. In addition to these variables the HLA repertoire of a patient has also been important in immune responses to therapeutic proteins [7,21]. We had therefore previously HLA-typed 1,000 patients who are included in the MLOF data set and showed that the HLA repertoire of a patient can be either a risk factor for, or protective against inhibitor development [7]. In addition to the HLA repertoire of a patient, it has also been previously shown that patients with large deletions and nonsense variants have a higher risk of inhibitor formation [3]. These studies have prompted the hypothesis that the number of FVIII derived foreign peptides that can potentially engage with HLA-II proteins could be a marker for inhibitor risk [5]. We quantified the number of foreign peptides for each patient by aligning the sequence of the endogenous FVIII and the infused FVIII drug. Furthermore, computational, and *in vitro* studies have indicated that the affinity with which FVIII-derived foreign peptides bind to an individual patients HLA II variants is also a variable associated with inhibitor development [22–24]. Using the subset of 1,000 HLA typed patients we also computed the affinity of foreign peptides for that patient's HLA-II repertoire. The completed set of variables associated with each patient which includes those provided by MLOF as well as those computed by us is provided in Table 2.

In this study we used unbiased AI-based approaches to simultaneously interrogate all variables to identify those associated with an inhibitor positive status. We used the best available ML classification models for prediction of FVIII inhibition, and then applied XAI to

the top performing model to determine what variables are important for FVIII protein inhibition. The data were processed and trained with multiple ML models to identify top performing models using metrices for accuracy, precision, recall, and F1-score. The results, reported using SHAP, can be ranked. The results depicted in Fig. 4, rank variables on the basis of mean SHAP value (wherein a higher mean SHAP value is associated with a greater impact on the FVIII inhibitor prediction). The LGBM model was picked as the best performing model, where the top five variables for predicting inhibitor development based on SHAP values are: (i) baseline activity of the FVIII protein, (ii) mean affinity of all foreign peptides for HLA DRB 3, 4, & 5 alleles, (iii) the minimum affinity among all foreign peptides for HLA DRB1 alleles, (iv) mean affinity of all foreign peptides for HLA DRB1 alleles, and (v) *F8* mutation type. While the other input variables were all important in the prediction of the inhibitor development, the top five variables listed above were highlighted as having a larger impact.

The variables with high SHAP values are all directly or indirectly associated with the genetic defect in the *F8* gene. The variables with high SHAP values are all directly or indirectly associated with the genetic defect in the *F8* gene. (i) Baseline activity of the FVIII protein: Within our cohort patients with a missense mutation in the I gene have a significantly ($P = 2.09 \times 10^{-17}$) higher FVIII activity compared to patients with larger disruptions in the FVIII gene (e.g., deletion of multiple exons). (ii) mean affinity of all foreign peptides for HLA DRB 3, 4, & 5 alleles, (iii) the minimum affinity among all foreign peptides for HLA DRB1 alleles, (iv) mean affinity of all foreign peptides for HLA DRB1 alleles: Foreign peptide-HLA affinities are determined by (a) identifying mismatch between the infused FVIII drug and the endogenous FVIII sequence and, (b) estimating in silico, the affinities of all foreign (i.e., mismatched) peptides and the patient's HLA repertoire. Foreign peptides identified thus depend on the genetic variants found in the individual patients *F8* gene. (v) *F8* mutation type: This variable directly evaluated the genetic defect in the patient's *F8* gene. These findings are consistent with our previous work with a small cohort of 25 hemophilia A patient [5], which showed that foreign-peptide HLA-II binding affinity was a better predictor of inhibitors than the number of foreign peptides *per se*. In the work reported here we find that foreign-peptide HLA-II binding affinities have high SHAP values for predicting inhibitors.

With respect to the underlying immunology the presentation of foreign peptides to the immune system by antigen presenting cells is an initial necessary step in the subsequent development of antibodies. Peptides with high affinity for the HLA-II proteins would be presented in larger numbers by antigen presenting cells and thus be more likely to elicit an immune response. As each patient can generate multiple foreign FVIII derived peptides we used both the average peptide-HLA-II binding affinity and the minimum percentile score (strongest binder) as independent variables. Both measures provide a reasonable rationale for association with inhibitor development. It could be argued that even a single peptide with very high affinity to the patient's HLA repertoire could elicit an immune response. Similarly, a stronger average binding affinity implies that multiple peptides could bind with relatively high affinity. The SHAP analyses suggest that both foreign peptide mean and minimum MHC-II binding affinities have similar values. *F8* mutation type is the final variable among those with the five highest SHAP values.

Among the other variables reported in the literature as being associated with inhibitor development in HA, race features prominently [4,8,9]. Consistent with previous studies we found that in our cohort, that >35% of black patients and <25% of white patients developed inhibitors. A larger dataset can be utilized to independently determine how the genetic variables we identify distribute between black and white populations. However, per the AI-based model, race does not have a very high SHAP value for predicting inhibitor development in a patient (Fig. 4 (A)). Similarly, inhibitor development in patients with the Intron-22-Inversion has been the subject of some considerable study as it is the most common gene defect in hemophilia A patients. The *F8* genotype is an important determinant of inhibitor development in hemophilia A patients [3]. Compared to patients with intron 22 inversion, the risk of inhibitors in patients with large deletions and nonsense variants was higher. Mutation type also shows a high SHAP value in our AI based model.

The study presented here has some limitations. Our results are based purely on an unbiased AI approach, and not a statistical approach. Statistical approaches are based on a hypothesis and test specific hypotheses. The AI based approach we have adopted in this work is unbiased (or "hypothesis free") and seeks to identify variables associated with a particular outcome (in this case inhibitor development) from the dataset in its entirety. Variables such as race and benign *F8* variants were both input variables but specific interactions between them and other variables were not reported in this study as we aimed at the main objective of identifying specifically which variables were more prevalent for the development of FVIII inhibitors. Even though AI/ML systems provide a newer and novel insights for the treatment of Hemophilia A, there are limitations that need to be noted. For example, the lack of proper datasets that include all the relevant medical and biological data is well noted in literature. Especially from a ML perspective, datasets that are large enough to train and test a model with good performance is a limiting obstacle. Therefore, having a larger dataset would offset some of the limitations from a ML perspective such as achieving 100% performance across all metrices. The tradeoff between explainability and performance is also a major limitation as black-box models like artificial neural networks are more difficult to explain. Therefore, the current study's interpretability is limited because it did not restrict the research to the most clinically significant phenotypes, which could have introduced bias.

The genomic data from this cohort is very rich, especially when used in conjunction with the HLA typing carried out previously by us. This dataset can thus be interrogated using diverse approaches. The set of variables available also vary with respect to individual subjects. Some approaches are more tolerant to such gaps while others are not. Thus, some analyses may benefit from a smaller but better curated data set. We would also like to note that there are other potential risk factors for inhibitor development, including polymorphisms in the TNFA, IL10, and CTLA-4 genes [25–27]. Since we have the genomic data from the volunteers, one future research direction for our group is the application of XAI to test hypotheses regarding these genes and their relevancy to inhibitor development, in conjunction with our findings here.

Taken together, the AI-based analysis of the MLOF dataset demonstrates that patient-related genetic factors are predictive for inhibitor development in hemophilia A patients. Importantly, variables that incorporate information about variants in the *F8* gene as

well as the HLA repertoire of the patient are better predictors. These findings can be translated into algorithms that can help physicians better personalize *F8* replacement therapies and achieve better clinical outcomes.

5. Author contributions

ZES conceived and designed the research. AR, CK, JO performed the research, AR, CK, JO, ONY, HY analyzed and interpreted the data, and AR, CK, JO, ONY, HY and ZES wrote the paper. All authors contributed to the article and approved the submitted version.

Funding

This research was supported by funds from the Hemostasis Branch/Division of Plasma Protein Therapeutics/Office of Tissues and Advanced Therapies/Center for Biologics Evaluation and Research of the U.S. Food and Drug Administration and in part by an appointment to the Research Participation Program at the Center for Biologics Evaluation and Research administered by the Oak Ridge Institute for Science and Education through an interagency agreement between the U.S. Department of Energy and the U.S. Food and Drug Administration. The MLOF program was developed as a partnership between NHF, ATHN, Bloodworks Northwest, and Bioverativ and supported financially by Bioverativ, NHF, Bloodworks Northwest, and ATHN.

Author contribution statement

Atul Rawal, Christopher Kidchob, Jiayi Ou: Performed the experiments; Analyzed and interpreted the data; Wrote the paper.
Osman N. Yogurtcu, Hong Yang: Analyzed and interpreted the data; Wrote the paper.
Zuben E. Sauna: Conceived and designed the experiments; Wrote the paper.

Data availability statement

Data already available in a public repository.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This study used the computational resources of the High-Performance Computing clusters at the Food and Drug Administration, Center for Devices and Radiological Health.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.heliyon.2023.e16331.

References

[1] P.M. Mannucci, E.G.D. Tuddenham, The hemophilias — from royal genes to gene therapy, N. Engl. J. Med. 344 (23) (2001) 1773–1779.
[2] P.H. Cygan, P.A. Kouides, Regulation and importance of factor VIII levels in hemophilia A carriers, Curr. Opin. Hematol. 28 (5) (2021) 315–322.
[3] S.C. Gouw, et al., F8 gene mutation type and inhibitor development in patients with severe hemophilia A: systematic review and meta-analysis, Blood 119 (12) (2012) 2922–2934.
[4] K.R. Viel, et al., Inhibitors of factor VIII in black patients with hemophilia, N. Engl. J. Med. 360 (16) (2009) 1618–1627.
[5] G.S. Pandey, et al., Endogenous factor VIII synthesis from the intron 22-inverted F8 locus may modulate the immunogenicity of replacement therapy for hemophilia A, Nat. Med. 19 (10) (2013) 1318–1324.
[6] G.S. Pandey, et al., Polymorphisms in the F8 gene and MHC-II variants as risk factors for the development of inhibitory anti-factor VIII antibodies during the treatment of hemophilia a: a computational assessment, PLoS Comput. Biol. 9 (5) (2013), e1003066.
[7] J.R. McGill, V.L. Simhadri, Z.E. Sauna, HLA variants and inhibitor development in hemophilia A: a retrospective case-controlled study using the ATHNdataset, Front. Med. 8 (2021), 663396.
[8] J. Addiego Jr., Increased frequency of inhibitors in African American hemophilia A patients, Blood 84 (1) (1994) 239a.
[9] L.M. Aledort, D.M. Dimichele, Inhibitors occur more frequently in African-American and Latino haemophiliacs, Haemophilia 4 (1) (1998) 68.
[10] J.I. Lorenzo, et al., Incidence of factor VIII inhibitors in severe haemophilia: the importance of patient age, Br. J. Haematol. 113 (3) (2001) 600–603.
[11] F. Peyvandi, et al., A randomized trial of factor VIII and neutralizing antibodies in hemophilia A, N. Engl. J. Med. 374 (21) (2016) 2054–2064.
[12] S.C. Gouw, et al., Factor VIII products and inhibitor development in severe hemophilia A, N. Engl. J. Med. 368 (3) (2013) 231–239.
[13] H.S.R. Rajula, et al., Comparison of conventional statistical methods with machine learning in medicine: diagnosis, drug development, and treatment, Medicina (Kaunas) 56 (9) (2020).
[14] W. Jankowski, et al., Peptides identified on monocyte-derived dendritic cells: a marker for clinical immunogenicity to FVIII products, Blood Adv. 3 (9) (2019) 1429–1440.

[15] C.K. Kasper, et al., A more uniform measurement of factor VIII inhibitors, Thromb. Haemostasis 34 (6) (1975) 869–872.
[16] B. Verbruggen, et al., The Nijmegen modification of the Bethesda assay for factor VIII: C inhibitors: improved specificity and reliability, Thromb. Haemostasis 73 (2) (1995) 247–251.
[17] C. Miller, et al., Validation of Nijmegen–Bethesda assay modifications to allow inhibitor measurement during replacement therapy and facilitate inhibitor surveillance, J. Thromb. Haemostasis 10 (6) (2012) 1055–1061.
[18] A. Srivastava, et al., WFH guidelines for the management of hemophilia, 3rd edition, Haemophilia 26 (6) (2020) 1–158.
[19] Z.E. Sauna, et al., Evaluating and mitigating the immunogenicity of therapeutic proteins, Trends Biotechnol. 36 (10) (2018) 1068–1084.
[20] H.A.D. Lagasse, Q. McCormick, Z.E. Sauna, Secondary failure: immune responses to approved protein therapeutics, Trends Mol. Med. 27 (11) (2021) 1074–1083.
[21] Z.E. Sauna, et al., Evaluating and mitigating the immunogenicity of therapeutic proteins, Trends Biotechnol. 36 (10) (2018) 1068–1084.
[22] C.L. Kempton, A.B. Payne, HLA-DRB1-factor VIII binding is a risk factor for inhibitor development in nonsevere hemophilia: a case-control study, Blood Adv 2 (14) (2018) 1750–1755.
[23] A.J. Shepherd, et al., A large-scale computational study of inhibitor risk in non-severe haemophilia A, Br. J. Haematol. 168 (3) (2015) 413–420.
[24] C. Yanover, et al., Pharmacogenetics and the immunogenicity of protein therapeutics, Nat. Biotechnol. 29 (10) (2011) 870–873.
[25] J. Astermark, et al., Polymorphisms in the TNFA gene and the risk of inhibitor development in patients with hemophilia A, Blood 108 (12) (2006) 3739–3745.
[26] J. Astermark, et al., Polymorphisms in the IL10 but not in the IL1beta and IL4 genes are associated with inhibitor development in patients with hemophilia A, Blood 107 (8) (2006) 3167–3172.
[27] J. Astermark, et al., Polymorphisms in the CTLA-4 gene and inhibitor development in patients with severe hemophilia A, J. Thromb. Haemostasis 5 (2) (2007) 263–265.