








## Article

# High-dimensional analysis of single-cell flow cytometry data predicts relapse in Childhood Acute Lymphoblastic Leukemia

Salvador Chulián <sup>1,2,†,‡</sup> , Álvaro Martínez-Rubio <sup>1,2,‡</sup> , Víctor M. Pérez-García <sup>3,4,5,\*</sup> , María Rosa <sup>1,2</sup> , Cristina Blázquez Goñi <sup>6</sup>, Juan Francisco Rodríguez Gutiérrez <sup>6</sup>, Lourdes Hermosín-Ramos <sup>6</sup>, Águeda Molinos Quintana <sup>7</sup>, Teresa Caballero-Velázquez <sup>7</sup> , Manuel Ramírez-Orellana <sup>8</sup> , Ana Castillo Robleda <sup>8</sup>, Juan Luis Fernández-Martínez <sup>9</sup> 

<sup>1</sup>Department of Mathematics, Universidad de Cádiz, Puerto Real, Cádiz, Spain

<sup>2</sup>Biomedical Research and Innovation Institute of Cádiz (INiBICA), Hospital Universitario Puerta del Mar Cádiz, Spain

<sup>3</sup>Department of Mathematics, Mathematical Oncology Laboratory (MOLAB), Universidad de Castilla-La Mancha, Ciudad Real, Spain

<sup>4</sup>Instituto de Matemática Aplicada a la Ciencia y la Ingeniería (IMACI), Universidad de Castilla-La Mancha, Ciudad Real, Spain

<sup>5</sup>ETSI Industriales, Universidad de Castilla-La Mancha, Ciudad Real, Spain

<sup>6</sup>Department of Paediatric Hematology and Oncology, Hospital de Jerez Cádiz, Spain

<sup>7</sup>Department of Paediatric Haematology and Oncology, Hospital Virgen del Rocío, Sevilla, Spain

<sup>8</sup>Department of Paediatric Haematology and Oncology, Hospital Infantil Universitario Niño Jesús, Instituto Investigación Sanitaria La Princesa, Madrid, Spain

<sup>9</sup>Department of Mathematics, Group of Inverse Problems, Optimization and Machine Learning, University of Oviedo, Oviedo, Spain

† Current address: Department of Mathematics, Universidad de Cádiz, Puerto Real, Cádiz, Spain

‡ These authors contributed equally to this work.

**Abstract:** Artificial intelligence methods may help in unveliling information hidden in high-dimensional oncological data. Flow cytometry studies of haematological malignancies provide quantitative data with the potential to be used for the construction of response biomarkers. Many computational methods from the bioinformatics toolbox can be applied to these data but have not been exploited in their full potential in leukaemias, specifically for the case of childhood B-cell acute lymphoblastic leukemia. In this paper we analysed flow cytometry data obtained on diagnosis from 54 paediatric B-cell acute lymphoblastic leukemia patients from two local institutions. We constructed classifiers based on the Fisher's Ratio to quantify differences in expression levels of immunophenotypical markers between patients with relapsing and non-relapsing disease. The distribution of the marker CD38 was found and validated to have a strong discriminating power between both patient cohorts, thus providing a classifier.

**Keywords:** Acute Lymphoblastic Leukaemia; Flow Cytometry Data; Fisher's Ratio; CD38; mathematical oncology; response biomarkers; personalized medicine

## 1. Introduction

Acute Lymphoblastic Leukaemia (ALL) is the most common childhood cancer, accounting for 40% of all paediatric neoplasias [1]. This disease is characterised by the abnormal growth of immature lymphocytes in the bone marrow (BM). ALLs are classified as B- or T-ALL depending on the lineage of the cells of origin of the malignancy [2]. The former comprises the majority of cases in children and has better prognosis than the later. The clinical manifestations of B-ALL are the result of the invasion of the bone marrow, having more than 25% of blasts or immature lymphocytes, which leads to a shortage of healthy haematological cells. Progresses in diagnosis, risk assessment and therapeutics

have increased survival rates to around 80% [3]. The prognosis of relapsing patients is substantially worse. The early identification of relapsing patients is of high clinical interest since it could allow to use immunotherapy or bone marrow transplant where appropriate, as a first line treatment [4].

Flow cytometry is widely used in the diagnosis and follow up of B-ALL [5]. This technique allows the measurement of the surface expression levels of selected proteins for individual cells. Since each cell development state is characterised by a set of these markers, flow cytometry allows to classify the different cell populations within the BM in comparison with the normal BM and assist with the disease diagnosis [6]. Typical diagnostic flow cytometry studies interrogate between  $10^5$  and  $10^6$  cells. The outcome of the analysis is a dataset with surface protein expressions, complexity and size at the single-cell level. Flow cytometers used in clinical contexts can detect between 4 and 18 markers [7]. The size of the flow-cytometry datasets is increasing as technology progresses, and at some point the manual analysis carried out nowadays by cytometrists will no longer be feasible [8].

Combinations of single cell cytometric data and bioinformatics pipelines have been recently used for biomarker discovery in lymphoma [11], renal cell carcinoma [12], melanoma [13] or lung adenocarcinoma [14]. In childhood B-cell ALL, there are studies on automated follow up [15] and relapse prediction, using either clinical characteristics [16] or mass cytometry data [17]. Having biomarkers of response to current chemotherapeutical protocols on diagnosis is of high interest in order to consider alternative therapeutic options, such as bone marrow transplants or CAR-T cells.

The aim of this study was to find a collection of surface proteins showing significant differences in expression levels between relapsed and non-relapsed patients on diagnosis. To do we took advantage of the high-dimensionality of flow cytometry data and a multicentre database of patients. To do so we performed several tasks. First, data had to be preprocessed to solve problems like missing data values or data imbalance. Next, we had to find a subset of relevant features to be used for classification, what we addressed using Fisher's linear discriminant analysis. Finally, a mathematical model using those features classifying patients on diagnosis was developed and validated.

## 2. Materials and Methods

### 2.1. Patients

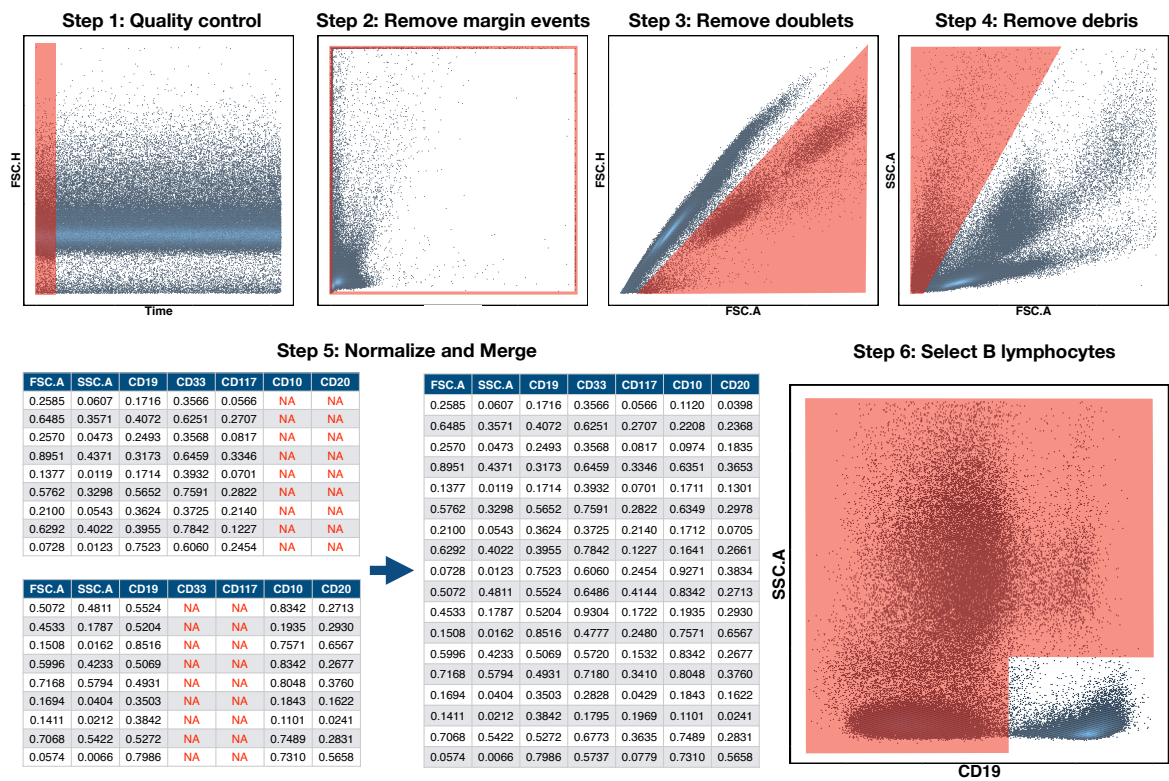
A retrospective study was designed in accordance with the Declaration of Helsinki, and the protocol was approved by the institutional review board (IRB) of the two participating local institutions (LLAMAT Project, 2018).

Inclusion criteria for the study were ALL diagnosis between February 2009 and October 2017, age over 1 year and less than 19 years, and availability of bone marrow flow cytometry data. A total of 105 patients satisfied the inclusion criteria. Exclusion criteria were availability of Flow Cytometry files FCS below 3.0, patients without a minimum of 15 immunophenotypic (IPT) markers in common with others in the dataset, and insufficient follow-up for non-relapsing patients, i.e. patients without relapse but with less than three years after no refractory values for minimal residual disease were found. Finally, 54 patients diagnosed in two of the local institutions were retained for further analysis. Dataset 1 included 28 non-relapsed and 8 relapsed patients, while dataset 2 included 13 non-relapsed and 5 relapsed patients.

### 2.2. Flow cytometer machines and antibodies

Marker expression was obtained on FACSCanto II flow cytometers, in accordance with the manufacturer's specifications for sample preparation. Final samples were stained using an 8-colour panel with six fluorochrome-conjugated antibodies.

FCS 3.0 files contained information on forward scatter (FSC) (interpreted as size), side scatter (SSC) light (interpreted as complexity) and monoclonal antibodies used routinely in diagnosis. The markers used included B-cell (CD19, CD10, CD20, CD22, CD24, IgM, CD66c, CD79a, kappa, lambda,



**Figure 1. Preprocessing pipeline.** Preprocessing was carried out in six steps. The first four were performed in FlowJo and consisted in the removal of abnormal acquisitions (quality control), margin events, doublets and debris. Files were then imported into R in step 5, and for each patient all tubes or aliquots were merged into a single file by means of nearest-neighbour imputation. Finally, in step 6, the CD19<sup>+</sup> population (B cells) was selected for further analysis.

etc.) and T-cell-related (CD7, cyCD3) IPT markers, markers related to the myeloid lineage (CD9, CD13, CD33, CD123), and some general ones (CD15, CD34, CD38, CD45, CD58, CD71, HLA-DR).

2.3. Preprocessing

Files were first imported into FlowJo (Becton Dickinson, 10.6.1) and FACSDiva (Becton Dickinson, 8.0.1) and inspected manually. Quality control was performed and margin events, debris, dead cells and doublets were removed, as shown in Figure 1 steps 1-4. Files were then further processed in R (3.6.0) and RStudio (1.2.1335). This software, in conjunction with Bioconductor (3.11) provides packages and methods for analysis of flow cytometry data. Tubes were compensated by means of the spillover matrix included in each file and then transformed with the Logicle transformation [18] included in the flowCore package (2.0.1) [19] with parameters  $w = 0.75$ ,  $t = 262144$ ,  $m = 4.5$  and  $a = 0$ . Our next step was to bring into a single file the information contained in each of the patient's tubes. Since each tube contains marker intensity for different markers and cells, the full set of 20 markers was not available for any of the cells, as shown in Figure 1 step 5. This posed a problem of missing data imputation, that is addressed in different ways in the context of flow cytometry [20–23]. We followed the methodology described in [20], which consist of nearest-neighbour imputation using the common or backbone markers in all aliquots. The result of this process was a set of 38 files, one per patient, containing complete information of the 20 IPT markers selected for the analysis. After this step,  $10^5$  events were randomly sampled from each file in order to have the same number of cells for each patient.

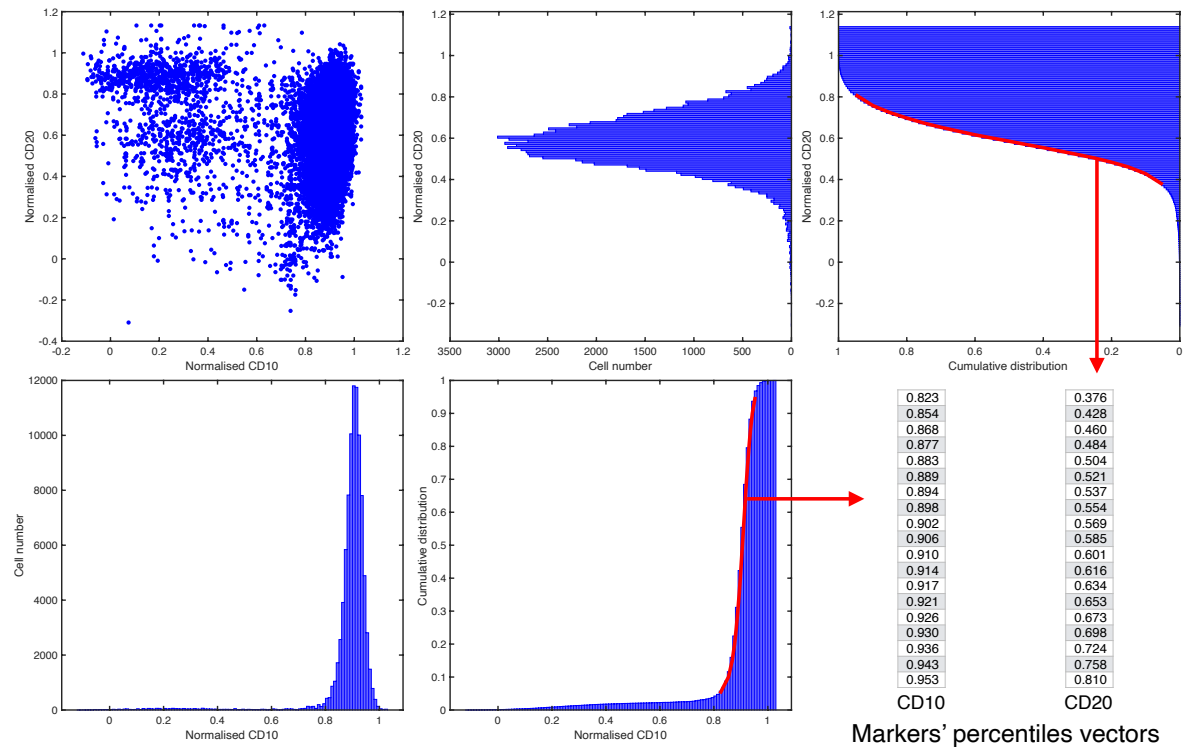
Since data of multicenter retrospective studies can be affected by batch effects and technical variations across time and centre, we performed a normalization based on a modified min-max

transformation. This transformation brings all data points to the range  $[0, 1]$  but it is sensitive to outliers. Instead of selecting the maximum and minimum values, we chose quantiles 0.05 and 0.95 and applied the transformation:

$$x' = \frac{x - x_{q0.05}}{x_{q0.95} - x_{q0.05}}, \quad (1)$$

where  $x_{q0.05}$  is the 5<sup>th</sup> percentile and  $x_{q0.95}$  is the 95<sup>th</sup> percentile. Finally, we used the common B-cell antigen CD19 to select the B-cell subpopulation, as shown in Figure 1 step 6. Files were imported in MATLAB (Mathworks, R\_2020a) via the *fca\_readfcs* function [24].

#### 2.4. Fisher's linear discriminant for relapse prediction



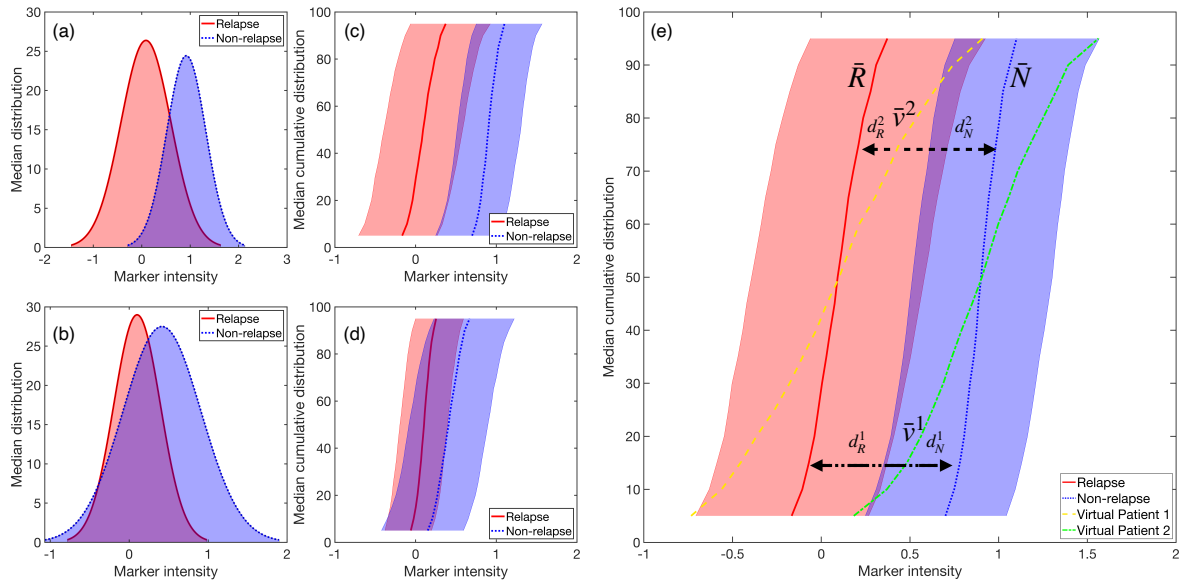
**Figure 2. Percentile vector construction.** (a) Scatter plot of a patient  $i$  for two normalised parameters,  $k_1=CD10$  and  $k_2=CD20$ . (b1) and (b2). Histograms cell count of, respectively,  $k_1=CD10$  and  $k_2=CD20$ . (c1) and (c2). Cumulative distribution of markers  $k_1=CD10$  and  $k_2=CD20$ , respectively. In red, percentiles curve from 5<sup>th</sup> to 95<sup>th</sup> percentile. (d) Each percentile curve for each patient  $i$  and marker  $k$  results in a vector  $P_{ij}$ , where  $j$  represents each percentile chosen.

Let us consider then  $P_{ij} \in \mathbb{R}^p$  as vectors obtained for each patient  $i$  and each common feature  $j$ , for  $i = 1, \dots, n$  patients and  $j = 1, \dots, m$  IPT markers. Thus, for each patient  $i$ , this results in a matrix  $M \in \mathbb{R}^{m \times p}$  of the  $p$  percentiles from all IPT markers  $m$ , as shown in Figure 2. Let us define the general Fisher's Ratio (FR) Matrix  $FR \in \mathbb{R}^{m \times p}$  [25], where

$$FR_{jk} = \frac{(\mu_{R_{jk}} - \mu_{N_{jk}})^2}{\sigma_{R_{jk}}^2 + \sigma_{N_{jk}}^2}, \quad (2)$$

for each IPT marker  $j$  in each percentile  $k$ , for  $j = 1, \dots, m$  and  $k = 1, \dots, p$ . In this case,  $\mu_{R_{jk}}$  and  $\mu_{N_{jk}}$  are the median percentiles  $k$  for the IPT marker's distribution  $j$  in each class of patients.  $R$  stands for relapsing patients, while  $N$  refers to non-relapsing ones. Parameters  $\sigma_{R_{jk}}$  and  $\sigma_{N_{jk}}$  are the standard deviation measures within the classes.





**Figure 3.** Example of synthetic IPT markers distributions. Mean distribution of a marker with respectively (a) high and (b) low Fisher's Ratio, with (c) and (d) their respective cumulative distribution of the median  $\pm$  the standard deviation values. (e) Median cumulative distribution of the two sets of patients for a marker with high Fisher's Ratio. In solid red line, median cumulative distribution of relapsed patients  $\bar{R}$  and in blue dotted line for the non-relapse ones. In yellow dashed line and green dashed dotted line the median cumulative distribution for the marker  $\bar{v}^i$  was represented for two different virtual patients  $i$ . The distances to each set median,  $d_R^i$  and  $d_N^i$ , are represented with black headed arrows, with dashed lines for Patient 1 and dashed dotted lines for Patient 2. In this example, Patient 1 would be considered as a relapsed patient, while Patient 2 would belong in the non-relapsed set.

To construct a classifier, we can select the highest  $FR_{jk}$  for  $j = 1, \dots, m$  and  $k = 1, \dots, p$ , thus obtaining percentiles from several IPT markers with lowest deviation and highest difference in median between each subset. Thus we would obtain a general discriminant classifier of a maximum of  $m^* \leq m$  markers and a maximum associated discriminant percentiles  $p^* \leq p$ .

In order to classify the patients, let us consider a certain IPT marker  $j$  and percentiles  $k$  for each class of patients. We can associate it then to an specific central measure  $\mu_{R_{jk}}$  or  $\mu_{N_{jk}}$  and dispersion measure  $\sigma_{R_{jk}}$  or  $\sigma_{N_{jk}}$ . Thus we set two control points

$$\begin{aligned}\bar{R}_{jk} &= \frac{\mu_{R_{jk}}}{\sigma_{R_{jk}}}, \\ \bar{N}_{jk} &= \frac{\mu_{N_{jk}}}{\sigma_{N_{jk}}},\end{aligned}\tag{3}$$

of the central measure of the IPT marker for each class of patients, normalised by the dispersion measure. If we consider now a patient not assigned to any set and  $\bar{v}_{jk}$  as the vector containing the percentiles  $k$  of a IPT marker  $j$ , we can compute a control point

$$\bar{P}_{jk} = \frac{\bar{v}_{jk}}{\left(\frac{\sigma_{R_{jk}} + \sigma_{N_{jk}}}{2}\right)}.\tag{4}$$

This point is normalised by the mean of both dispersion measures, as we consider  $\bar{P}$  as a non-assigned patient control point. Now, we can use a distance function  $d : \mathbb{R}^k \times \mathbb{R}^k \rightarrow [0, \infty)$  to measure the separation between the new patient  $\bar{P}_{jk}$  and the control points  $\bar{R}_{jk}$  and  $\bar{N}_{jk}$  (Figure 3).

For each IPT marker, we construct a probability measure for each IPT marker and percentile  $\mathcal{P}$  as

$$\begin{aligned}\mathcal{P}(\bar{P}_{jk} \in R) &= \frac{d(\bar{P}_{jk}, \bar{R}_{jk})}{d(\bar{P}_{jk}, \bar{R}_{jk}) + d(\bar{P}_{jk}, \bar{N}_{jk})}, \\ \mathcal{P}(\bar{P}_{jk} \in N) &= \frac{d(\bar{P}_{jk}, \bar{N}_{jk})}{d(\bar{P}_{jk}, \bar{R}_{jk}) + d(\bar{P}_{jk}, \bar{N}_{jk})}.\end{aligned}\quad (5)$$

The mean of the probability measures for all the IPT markers selected for each patient may allow us to classify the patient in the relapsing or non-relapsing classes.

### 2.5. Validation and feature relevance

To validate the classification algorithm, both K-fold, and leave-one-out cross-validation techniques were applied. The resulting performance of each model was obtained by averaging over 20 evaluations each K-Fold, and 20 for Leave-One-Out cross-validation (LOOCV). Each technique was repeated in each evaluation to fully cover each data set. For both cases, a minimum of one patient of each set was always in the training set.

For each validation technique, we constructed a classifier with the most significant IPT markers  $j$  according to the Fisher Ratio ( $FR_{jk} > 0.5$ ). We computed in each simulation a receiver operating characteristic (ROC) curve and its associated Area Under Curve (AUC). Accuracy was obtained as the number of correctly classified samples divided by the total number of classified samples. Along with these magnitudes we computed from each confusion matrices the sensitivity, specificity, positive predictive value (PPV) and negative predictive value (NPV).

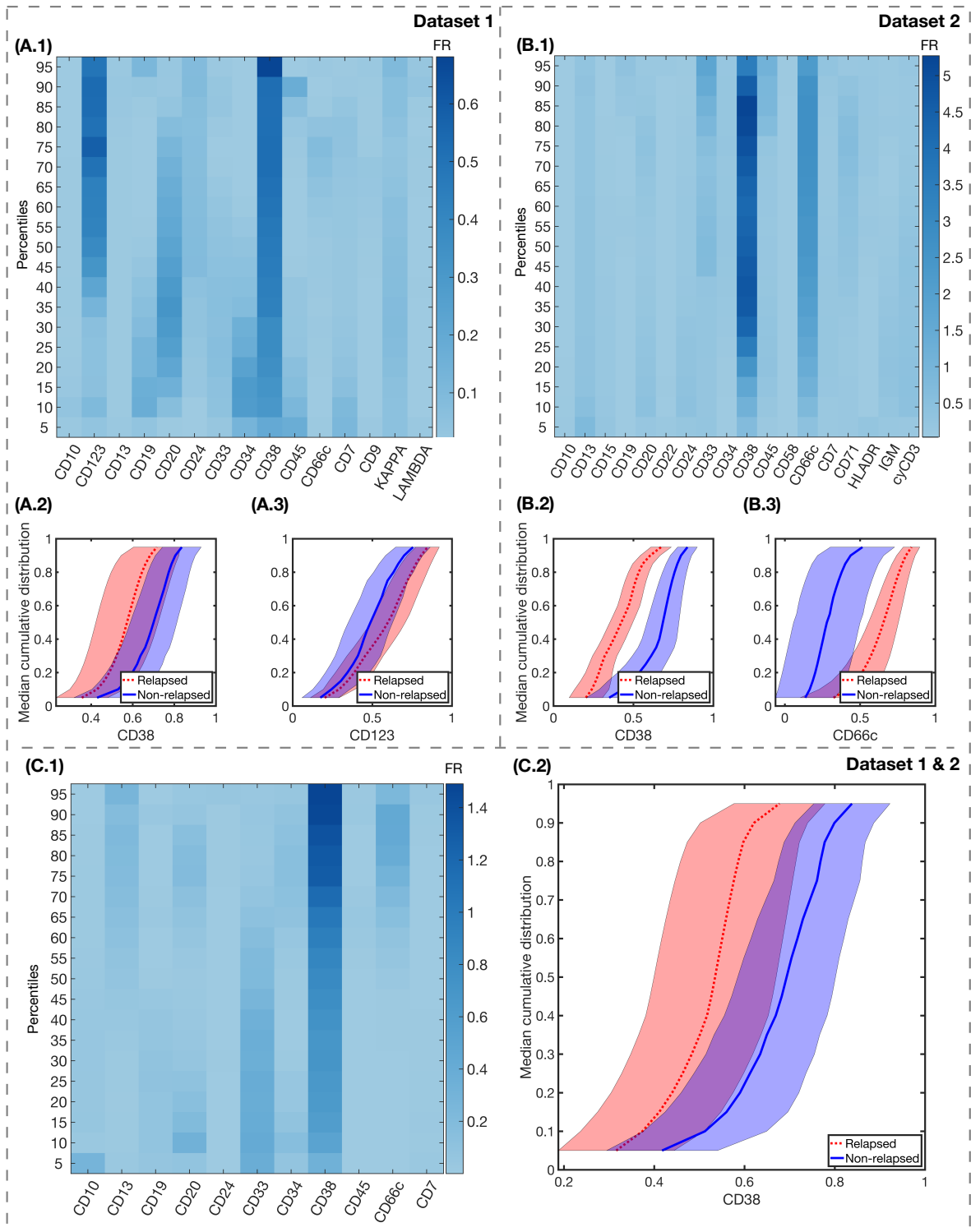
To choose the IPT markers with better prognostic value, we performed a Monte-Carlo based train-test split of the whole set of patients. We ran 100 simulations where each class of patients was divided into a 75% training and 25% test sets. We constructed a classifier for each splitting as described and then evaluated its accuracy. Once an accuracy threshold was fixed, we computed the frequency of every marker for the set of classifiers that were above that threshold. This calculation was performed for different values of the accuracy threshold.

Finally, we resorted to another method in order to compare the results. We performed 100 Random Forest classifications with 50 trees each and a 75-25 split of patients and then recorded the out-of-bag error and the permutation feature relevance.

## 3. Results

**CD38 distribution differed significantly between relapsing and non-relapsing patients.** We examined for Datasets 1 and 2 and for both of them combined those IPT markers with higher FR. Results are shown in Figure 4, as well as the median cumulative distribution of the arising markers.

For Dataset 1, CD38 FR was high in almost all percentiles, with  $FR_{jk} > 0.3$ , as seen in Figure 4 (A.1). IPT marker CD123 had high FR for the highest percentiles, with  $FR_{jk} > 0.3$  for  $k \in (50, 95)$ . For Dataset 2, the differences between FR were significantly higher, with  $FR_{jk} > 3.5$  in percentiles  $j \in (20, 95)$  for IPT marker CD38, and mean  $FR_{jk} > 2.5$  for IPT marker CD66c, as shown in Figure 4 (B.1.). For the combination of both datasets, only CD38 achieved a high FR with mean  $FR_{jk} > 0.9$  in all percentiles, as shown in Figure 4 (C.1.).



**Figure 4.** Fisher's Ratio Matrices for Dataset 1 (A.1), Dataset 2 (B.1), and both datasets combined (C.1). The common parameters within each dataset are represented in the x-axis, while in the y-axis we represent the percentiles of the median cumulative distribution. The intensity of the Fisher's Ratio for each percentile and markers are represented for each dataset in a colorbar for each chart. Median cumulative distributions and standard deviation bands of the IPT markers with higher FR, for relapsed (red, dotted lines) and non-relapsed (blue, solid lines) patients are represented in the following charts: for Dataset 1, CD38 (A.2) and CD123 (A.3); for Dataset 2, CD38 (B.2) and CD66c (B.3); and for both datasets combined, CD38 (C.2).

**Immunophenotypical markers CD38 and CD123, for dataset 1, and markers CD38 and CD66c, for dataset 2, predicted relapse after cross-validation.** K-fold and Leave-One-Out cross-validations were run in both directions: first, to know the most common IPT markers used in the training set for each simulation. This resulted in differences between median distribution differences of relapsed and non-relapsed patients again for the same sets of markers: CD38 and CD123 for Dataset 1 (Figures 4 (A.2.) and (A.3.)), CD38 and CD66c for Dataset 2 (Figures 4 (B.2.) and (B.3.)), and finally, only CD38 for the combination of both datasets (Figure 4 (C.2.)). Secondly, cross-validation techniques were repeated considering only common IPT markers. The results are shown in Table 1. The maximal number of folds was determined by the number of relapsing patients (8 for Dataset 1, and 5 for Dataset 2).

	Method	Accuracy	Sensitivity	Specificity	PPV	NPV	AUC
Dataset 1	LOOCV	$0.75 \pm 0.04$	$0.74 \pm 0.05$	$0.76 \pm 0.05$	$0.76 \pm 0.04$	$0.75 \pm 0.04$	$0.76 \pm 0.02$
	2-Fold	$0.59 \pm 0.1$	$0.63 \pm 0.14$	$0.43 \pm 0.2$	$0.81 \pm 0.04$	$0.24 \pm 0.12$	$0.56 \pm 0.1$
	4-Fold	$0.62 \pm 0.07$	$0.63 \pm 0.1$	$0.58 \pm 0.12$	$0.85 \pm 0.03$	$0.3 \pm 0.06$	$0.65 \pm 0.06$
	6-Fold	$0.64 \pm 0.05$	$0.66 \pm 0.05$	$0.58 \pm 0.13$	$0.85 \pm 0.04$	$0.31 \pm 0.06$	$0.67 \pm 0.06$
	8-Fold	$0.7 \pm 0.04$	$0.7 \pm 0.04$	$0.71 \pm 0.06$	$0.9 \pm 0.02$	$0.39 \pm 0.04$	$0.72 \pm 0.03$
Dataset 2	LOOCV	$0.66 \pm 0.06$	$0.95 \pm 0.05$	$0.37 \pm 0.1$	$0.6 \pm 0.04$	$0.88 \pm 0.1$	$0.89 \pm 0.05$
	2-Fold	$0.72 \pm 0.07$	$0.95 \pm 0.06$	$0.13 \pm 0.22$	$0.74 \pm 0.05$	$0.42 \pm 0.41$	$0.68 \pm 0.16$
	4-Fold	$0.78 \pm 0.04$	$0.95 \pm 0.05$	$0.34 \pm 0.15$	$0.79 \pm 0.03$	$0.81 \pm 0.2$	$0.86 \pm 0.06$
Datasets 1 & 2	LOOCV	$0.69 \pm 0.05$	$0.62 \pm 0.09$	$0.75 \pm 0.09$	$0.72 \pm 0.07$	$0.67 \pm 0.05$	$0.78 \pm 0.04$
	2-Fold	$0.64 \pm 0.13$	$0.6 \pm 0.17$	$0.75 \pm 0.12$	$0.87 \pm 0.09$	$0.38 \pm 0.08$	$0.73 \pm 0.11$
	4-Fold	$0.69 \pm 0.01$	$0.67 \pm 0.02$	$0.77 \pm 0.01$	$0.91 \pm 0.01$	$0.41 \pm 0.01$	$0.77 \pm 0.04$
	6-Fold	$0.7 \pm 0.02$	$0.68 \pm 0.02$	$0.77 \pm 0.01$	$0.91 \pm 0.01$	$0.42 \pm 0.02$	$0.79 \pm 0.02$
	8-Fold	$0.7 \pm 0.01$	$0.68 \pm 0.02$	$0.77 \pm 0.01$	$0.91 \pm 0.01$	$0.42 \pm 0.02$	$0.79 \pm 0.02$
	10-Fold	$0.7 \pm 0.01$	$0.68 \pm 0.02$	$0.77 \pm 0.01$	$0.91 \pm 0.01$	$0.42 \pm 0.01$	$0.8 \pm 0.02$
	12-Fold	$0.69 \pm 0.01$	$0.67 \pm 0.02$	$0.77 \pm 0.01$	$0.91 \pm 0.01$	$0.41 \pm 0.01$	$0.79 \pm 0.01$

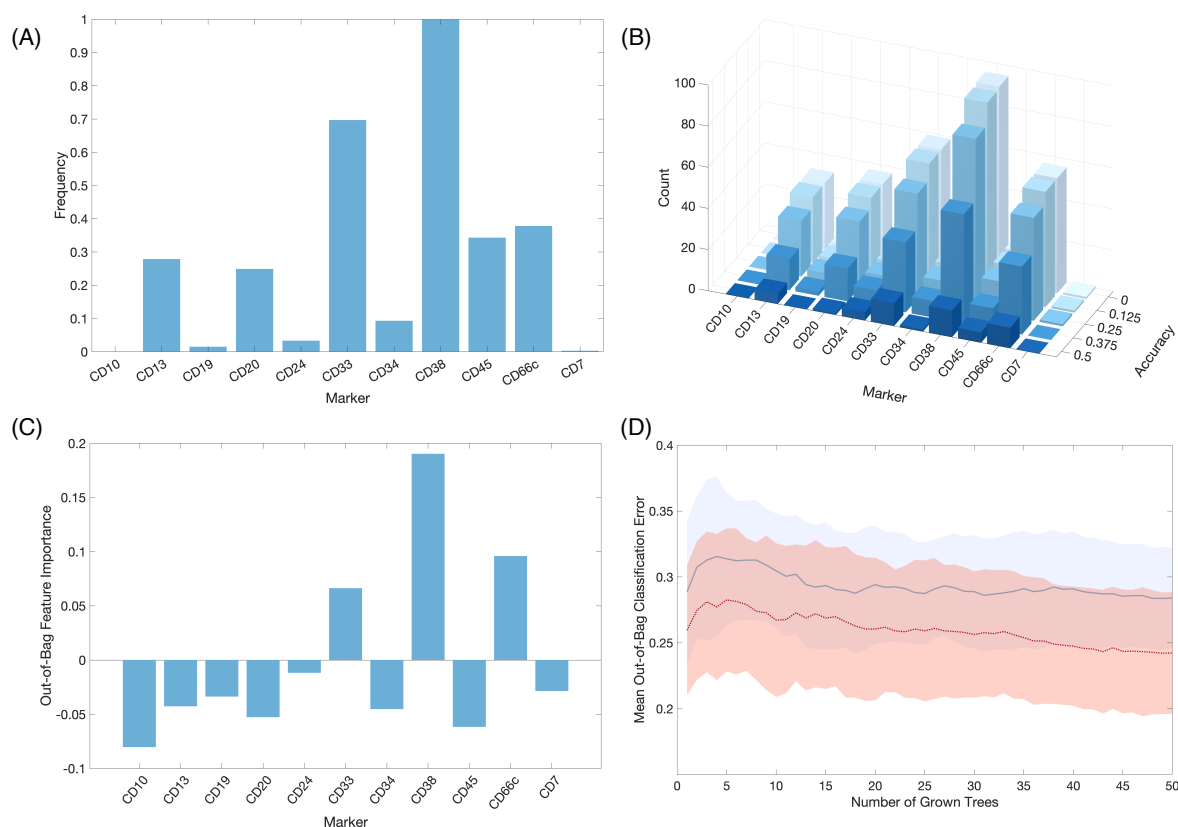
**Table 1.** Validated predictive performance of best classifiers CD38 and CD123 (Dataset 1), CD38 and CD66c (Dataset 2) and CD38 (Datasets 1& 2). PPV: Positive Predictive Value. NPV: Negative Predictive value. AUC: Area under curve.

**Train-test splitting revealed other markers with potential predictive value.** We tested the accuracy of the variables by splitting Dataset 1 and 2 combined into a training and test set with ratio 75:25. After 100 simulations, the frequency of the IPT markers used in the classifiers is shown in Figure 5(A). IPT marker CD38 arose again as the marker used in all classifiers, while CD33 was used on almost 70% of them. Having obtained the accuracy for the 100 classifiers, we count the number of IPT markers whose prediction accuracy is above a certain threshold, as shown in Figure 5(B). IPT markers CD13, CD24, CD33, CD38, CD45 and CD66c are those with an accuracy higher than 0.5%.

**Random-Forest analysis matched the results from the constructed classifiers.** Random Forest analysis after 100 simulations was considered, resulting in IPT markers CD33, CD38 and CD66c as those only with positive Out-of-bag Feature importance. However, after repeating the simulations only considering these markers, Out-Of Bag Classification Error was not significantly lower in comparison to the analysis with the whole set of IPT markers (mean out-of-bag error of 0.28 versus 0.31, respectively). Nevertheless, feature importance coincided with those markers with highest frequency in our previous classification.

#### 4. Discussion and conclusions

The unprecedented amount and complexity of clinical data that is available nowadays has resulted in the proliferation of bioinformatics pipelines and artificial intelligence algorithms to extract information from data. In flow cytometry, the routine analysis that is carried out by visualizing histograms and bidimensional plots is falling behind technical progress in the field [31]. Machine learning algorithms have the potential to speed up, automatize and reduce bias in conventional analyses, but also to complement the work done by the human operator [32]. Recent examples in haematology include leukocyte recognition, prediction of refractory Hodgkin lymphoma, minimal



**Figure 5.** Results for both datasets combined from train-test split and Random Forest analysis. **(A)** Frequency of the markers in all classifiers after 100 simulations of train-test splitting. **(B)** Histograms of the number of markers after establishing a threshold for the accuracy. **(C)** Out-of-bag feature importance of the markers after 100 simulations of Random Forest. **(D)** Mean and standard deviation bands of the Out-of-bag Classification Error in Random Forest analysis for the whole set of markers (blue, solid line) and for the set of markers with positive feature importance CD33, CD38 and CD66c (red, dotted line).

residual disease detection in acute myeloid leukaemia, risk stratification in multiple myeloma or predicting resistance in myelodysplastic syndrome [33]. In childhood B-cell ALL, machine learning has taken advantage of clinical data in order to predict either diagnosis [34] or relapse [16], with the work of Good et al including proteomics data for the latter purpose [17]. Reiter et al proposed a way to automatize Minimal residual disease follow-up [15].

Leaving aside accuracy and prediction reliability concerns, which we can expect to be solved or dampened as the scientific production continues, there are a number of issues that still hamper the integration of AI and the respective clinical context. As happens in general with the relationship between mathematics and medicine, researchers in both ends often speak a different language [35]. Many AI algorithms behave as a “black boxes”, providing an outcome directly from raw data and hindering a mechanistic interpretation of the underlying phenomena. For clinical use, it is highly desirable that the features uncovered by these algorithms can be interpretable and actionable. As Radakovich et. al. puts it, “Algorithms can only be as clinically meaningful as the outcomes that they are designed to predict” [36].

In this work, we designed an intuitive algorithm allowing to identify on diagnosis patients with potential or relapse versus those with no risk of relapse in B-cell childhood ALL. We used flow cytometry data obtained at diagnosis from two local institutions and based the analysis on two concepts that are already employed in this context; the intensity and range of surface markers expression and the frequency of cells within that range. We took those two factors into account by assigning each patient



and marker its percentile curve and then used the Fisher's ratio to look for meaningful differences between both groups of patients. That approach allowed us to construct a classifier based on this measure in order to assess the significance of the previously obtained differences. Given the small sample size, we used cross-validation routines to assess the validity of the Fisher's ratio-based measure. Despite the exploratory nature of the study, we were able to find some common trends in the data.

Firstly, we observed that Fisher's ratio displays differences in expression levels between relapsing and non relapsing patients. This was specially significant for the second dataset. Given that both datasets were pre-processed identically, the difference in the significance of the measure could be due to either sample size or different acquisition routines in either hospital. We expect to have a clearer understanding of this after increasing the number of patients in each dataset or the number of datasets as a whole. K-fold cross validation showed that, when restricting the analysis to the most important features according to the previously calculated Fisher's ratio, the algorithm was able to separate better between relapsing and non relapsing patients, always using only data available on diagnosis.

Measurements of performance yielded good values for this biomarker as measured by Accuracy and AUC. However, although specificity was high, we obtained a low negative predictive value, i.e. the algorithm underperformed when detecting relapses. This could be due to the intrinsic unbalance in the datasets, with only 25% of relapsed patients. The relevant information, nonetheless, was the agreement in the extraction of the most important features. This was later confirmed by the Monte-Carlo based and Random Forest feature importance computation. Both approaches agreed in this selection, specially when being more restrictive with the classification accuracy in the first one.

The most consistent result, in the different analyses and for both local institutions, was the association between a lower expression of CD38 marker and relapse. CD38 is a surface receptor present in a broad variety of immune cells. It is considered to be a cell activation marker and operates both as a receptor and an enzyme [37]. In the B cell compartment, both bone marrow precursors and terminally differentiated cells express CD38 [38]. In the context of haematological disease, high CD38 levels have been associated with worse prognosis in chronic lymphocytic leukemia [39]. Previous studies have suggested that CD38 is a suitable therapeutic target in both AML and ALL [40,41]. There has been some controversy concerning the existence of a CD34+/CD38- population of leukaemia initiating stem cells [42–45]. In B-ALL, the accumulated evidence indicates that lower levels of CD38 could be associated with worse outcome in terms of survival [46–49]. Our results aligned with this evidence, suggesting that a higher frequency of low CD38 expressing B cells could be an early indicator of relapse risk.

Other markers that were found to be relevant in this study were CD33 and CD66c. These two markers are normally expressed in cells of the myeloid lineage, and they have been linked to paediatric B-ALL in the context of myeloid antigen expressing B-cell malignancies. This refers to the fact that some malignant B cells can express markers from the myeloid line. CD66c is the most frequently observed aberrant myeloid antigen in B-cell ALL. Upon studying the correlation of the expression of this antigen with known prognostic factors, previous studies have found that CD66c is associated with BCR/ABL translocation, which has been shown to confer the worst prognosis [50–52]. Here, we found that patients relapsing were more prone to overexpression of this marker on diagnosis. With respect to CD33, there has been some controversy with respect to its prognostic value, but evidence suggests that the presence of high expressing CD33+ cells identifies patients with worse prognosis [53], contrary to the differences exhibited by percentile curves here.

Finally, the immunophenotypical marker CD123 was also highlighted by Fisher's ratio but only in dataset 1. Its importance could not be further assessed since it was not available in dataset 2. This marker was first described as a marker of acute myeloid leukemia stem cells. It was later shown to be uniformly expressed in B-ALL blasts, being proposed for the detection of minimal residual disease [54,55] and identified recently as a potential target for immunotherapies [56,57]. Interestingly, high expression of CD123 correlated with hyperdiploid karyotype, an indicator of favourable prognosis

in childhood B-cell ALL [58]. In our cohort we found a high proportion of CD123 expressing cells in relapsing patients.

While CD38 differences were present through the whole range of expression of the marker, that was not the case for CD33 and CD123. For those markers differences were observed only in the low expression region for the former and in the high expression region for the latter. The fact that there is less evidence for their prognostic value suggests that the method presented here leads to significant results if there is a constant difference in expression levels between both sets of patients. This is indeed a limitation of the study; by representing the expression as a percentile curve, we may miss information that can be clinically relevant and that refers not to the frequency of cells or intensity of expression, but to the presence or absence of a given subpopulation. In this regard, we already mentioned that a subpopulation of CD34+/CD38- cells could be associated with leukaemia initiating cells, and the same could happen for a restricted subpopulation of CD34+/CD38-/CD123+, this one agreeing with the results presented in this paper.

Another limitation of our analysis is the nature of the data, a recurrent concern in artificial intelligence in haematology [36]. Apart from having only 54 patients, the set of relapsing patients represented only 25% of the whole dataset and that unbalance could introduce biases in the analysis. In the future, as we increase the size of our dataset, it would be better to increase the number of patients to carry out a 50/50 analysis. Further, there is the issue of data variability, given that it was collected retrospectively, belonging to patients from different years and hospitals. This highlights the importance of the preprocessing routine, which is also amenable to improvements in order to ensure the comparability of the samples. These weaknesses provide future lines of work. While in the process of recruiting more patients and hospitals, efforts will be directed towards the automation of the preprocessing workflow and towards the combination with more complex analysis like dimensionality reduction, network analysis and clustering. Finally, this work could be complemented with the inclusion of other clinical data like cytogenetics and molecular biology information, also relevant in the prognostic assessment of haematological diseases.

Notwithstanding these limitations, this work adds to the growing field of artificial intelligence in haematology and specifically in B-cell childhood acute lymphoblastic leukaemia. We attempted to delineate differences in marker expression between patients who relapse from the disease and those that respond to treatment, obtaining results that are directly interpretable from the clinical point of view. The main result would be the underexpression of surface marker CD38 in patients experiencing relapse after the first-line chemotherapy treatment. This is very important knowledge since it could aid in therapy personalization by considering alternative therapies as upfront therapies for patients with high risk of relapse.

**Author Contributions:** Conceptualization, Salvador Chulián, Víctor Pérez-García, María Rosa and Juan Luis Fernández-Martínez; Data curation, Salvador Chulián, Álvaro Martínez-Rubio, María Rosa, Cristina Blázquez-Goñi, Juan Francisco Rodríguez Gutiérrez, Lourdes Hermosín-Ramos, Águeda Molinos-Quintana, Teresa Caballero-Velázquez, Manuel Ramírez-Orellana and Ana Castillo Robleda; Formal analysis, Salvador Chulián; Funding acquisition, Víctor Pérez-García and María Rosa; Investigation, Salvador Chulián, Álvaro Martínez-Rubio and María Rosa; Methodology, Salvador Chulián and Juan Luis Fernández-Martínez; Project administration, Víctor Pérez-García, María Rosa, Cristina Blázquez-Goñi, Águeda Molinos-Quintana and Manuel Ramírez-Orellana; Resources, Cristina Blázquez-Goñi, Juan Francisco Rodríguez Gutiérrez, Lourdes Hermosín-Ramos, Águeda Molinos-Quintana, Teresa Caballero-Velázquez, Manuel Ramírez-Orellana and Ana Castillo Robleda; Software, Salvador Chulián and Álvaro Martínez-Rubio; Supervision, Víctor Pérez-García, María Rosa and Juan Luis Fernández-Martínez; Validation, Salvador Chulián and Álvaro Martínez-Rubio; Visualization, Salvador Chulián and Álvaro Martínez-Rubio; Writing – original draft, Salvador Chulián and Álvaro Martínez-Rubio; Writing – review & editing, Salvador Chulián, Álvaro Martínez-Rubio, Víctor Pérez-García, María Rosa, Cristina Blázquez-Goñi, Juan Francisco Rodríguez Gutiérrez, Lourdes Hermosín-Ramos, Águeda Molinos-Quintana, Teresa Caballero-Velázquez, Manuel Ramírez-Orellana, Ana Castillo Robleda and Juan Luis Fernández-Martínez. All authors have read and agreed to the published version of the manuscript.

**Funding:** This work was partially supported by the Fundación Española para la Ciencia y la Tecnología [UCA PR214], the Asociación Pablo Ugarte (APU, Spain), Junta de Comunidades de Castilla-La Mancha [SBPLY/17/180501/000154], and Ministry of Science and Technology, Spain [PID2019-110895RB-I00].

**Acknowledgments:** We would like to acknowledge the research group of Junta de Andalucía (Spain) group FQM-201 and the Mathematical Oncology Laboratory Group (MôLAB) from the University of Castilla-La Mancha.

**Conflicts of Interest:** The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results.

## Abbreviations

The following abbreviations are used in this manuscript:

IPT	Immunophenotypic
FR	Fisher's Ratio
CD	Cluster of Differentiation
ALL	Acute Lymphoblastic Leukaemia
ROC	Receiver Operating Characteristic
AUC	Area Under Curve
LOOCV	Leave-One-Out cross-validation

## References

1. P. A. Pizzo, D. G. Poplack, Principles and practice of pediatric oncology, Lippincott Williams & Wilkins, 2015. doi:10.1016/j.suronc.2006.05.001.
2. T. Terwilliger, M. Abdul-Hay, Acute lymphoblastic leukemia: a comprehensive review and 2017 update, Blood cancer journal 7 (6) (2017) e577–e577. doi:10.1038/bcj.2017.53.
3. C.-H. Pui, J. J. Yang, S. P. Hunger, R. Pieters, M. Schrappe, A. Biondi, A. Vora, A. Baruchel, L. B. Silverman, K. Schmiegelow, G. Escherich, K. Horibe, Y. C.M. Benoit, S. Izraeli, A. E. J. Yeoh, D.-C. Liang, J. R. Downing, W. E. Evans, M. V. Relling, C. G. Mullighan, Childhood acute lymphoblastic leukemia: progress through collaboration, Journal of Clinical Oncology 33 (27) (2015) 2938. doi:10.1200/JCO.2014.59.1636.
4. D. Bhojwani, C.-H. Pui, Relapsed childhood acute lymphoblastic leukaemia, The lancet oncology 14 (6) (2013) e205–e217. doi:10.1016/S1470-2045(12)70580-6.
5. E. G. Weir, M. J. Borowitz, Flow cytometry in the diagnosis of acute leukemia, in: Seminars in hematology, Vol. 38, Elsevier, 2001, pp. 124–138. doi:10.1016/S0037-1963(01)90046-0.
6. J. J. M. van Dongen, L. Lhermitte, S. Böttcher, J. Almeida, V. H. J. van der Velden, J. Flores-Montero, A. Rawstron, V. Asnafi, Q. Lécrovisse, P. Lucio, E. Mejstrikova, T. Szczepański, T. Kalina, R. de Tute, M. Brüggemann, L. Sedek, M. Cullen, A. W. Langerak, A. Mendonça, E. Macintyre, M. Martin-Ayuso, O. Hrusak, M. B. Vidriales, A. Orfao, EuroFlow antibody panels for standardized n-dimensional flow cytometric immunophenotyping of normal, reactive and malignant leukocytes, Leukemia 26 (9) (2012) 1908–1975. doi:10.1038/leu.2012.120.
7. E. Lugli, M. Roederer, A. Cossarizza, Data analysis in flow cytometry: the future just started, Cytometry Part A 77 (7) (2010) 705–713. doi:10.1002/cyto.a.20901.
8. C. E. Pedreira, E. S. Costa, Q. Lécrovisse, J. J. van Dongen, A. Orfao, Overview of clinical flow cytometry data analysis: recent advances and future challenges, Trends in biotechnology 31 (7) (2013) 415–425. doi:10.1016/j.tibtech.2013.04.008.
9. N. Aghaeepour, G. Finak, H. Hoos, T. R. Mosmann, R. Brinkman, R. Gottardo, R. H. Scheuermann, Critical assessment of automated flow cytometry data analysis techniques, Nature Methods 10 (3) (2013) 228–238. doi:10.1038/nmeth0513-445c.
10. Y. Saeys, S. V. Gassen, B. N. Lambrecht, Computational flow cytometry: helping to make sense of high-dimensional immunology data, Nature Reviews Immunology 16 (7) (2016) 449–462. doi:10.1038/nri.2016.56.
11. H. Zare, A. Bashashati, R. Kridel, N. Aghaeepour, G. Haffari, J. M. Connors, R. D. Gascoyne, A. Gupta, R. R. Brinkman, A. P. Weng, Automated analysis of multidimensional flow cytometry data improves diagnostic accuracy between mantle cell lymphoma and small lymphocytic lymphoma, American journal of clinical pathology 137 (1) (2012) 75–85. doi:10.1309/AJCPMMLQ67YOMGEW.

12. S. Chevrier, J. H. Levine, V. R. T. Zanotelli, K. Silina, D. Schulz, M. Bacac, C. H. Ries, L. Ailles, M. A. S. Jewett, H. Moch, M. v. d. Broek, C. Beisel, M. B. Stadler, C. Gedye, B. Reis, D. Pe'er, B. Bodenmiller, An immune atlas of clear cell renal cell carcinoma, *Cell* 169 (4) (2017) 736–749. doi:10.1016/j.cell.2017.04.016.
13. C. Krieg, M. Nowicka, S. Guglietta, S. Schindler, F. J. Hartmann, L. M. Weber, R. Dummer, M. D. Robinson, M. P. Levesque, B. Becher, High-dimensional single-cell analysis predicts response to anti-pd-1 immunotherapy, *Nature medicine* 24 (2) (2018) 144. doi:10.1038/nm.4466.
14. Y. Lavin, S. Kobayashi, A. Leader, E.-a. D. Amir, N. Elefant, C. Bigenwald, R. Remark, R. Sweeney, C. D. Becker, J. H. Levine, K. Meinhof, A. Chow, S. Kim-Shulze, A. Wolf, C. Medaglia, H. Li, J. A. Rytlewski, R. O.Emerson, A. Solovyov, B. D. Greenbaum, C. Sanders, M. Vignali, M. B. Beasley, R. Flores, S. Gnjatich, D. Pe'er, A. Rahman, I. Amit, M. Merad, Innate immune landscape in early lung adenocarcinoma by paired single-cell analyses, *Cell* 169 (4) (2017) 750–765. doi:10.1016/j.cell.2017.04.014.
15. M. Reiter, M. Diem, A. Schumich, M. Maurer-Granofszky, L. Karawajew, G. J. Rossi, R. Ratei, S. Groeneveld-Krentz, O. E. Sajaroff, S. Suhendra, M. Kampel, N. M. Dworzak, Automated flow cytometric mrd assessment in childhood acute b- lymphoblastic leukemia using supervised machine learning, *Cytometry. Part A : the journal of the International Society for Analytical Cytology*. doi:10.1002/cyto.a.23852.
16. L. Pan, G. Liu, F. Lin, S. Zhong, H. Xia, X. Sun, H. Liang, Machine learning applications for prediction of relapse in childhood acute lymphoblastic leukemia, *Scientific reports* 7 (1) (2017) 1–9. doi:10.1038/s41598-017-07408-0.
17. Z. Good, J. Sarno, A. Jager, N. Samusik, N. Aghaeepour, E. F. Simonds, L. White, N. J. Lacayo, W. J. Fantl, G. Fazio, G. Gaipa, A. Biondi, R. Tibshirani, S. C. Bendall, G. P. Nolan, K. L. Davis, Single-cell developmental classification of b cell precursor acute lymphoblastic leukemia at diagnosis reveals predictors of relapse, *Nature medicine* 24 (4) (2018) 474. doi:10.1038/nm.4505.
18. D. R. Parks, M. Roederer, W. A. Moore, A new “logicle” display method avoids deceptive effects of logarithmic scaling for low signals and compensated data, *Cytometry Part A: The Journal of the International Society for Analytical Cytology* 69 (6) (2006) 541–551. doi:10.1002/cyto.a.20258.
19. F. Hahne, N. LeMeur, R. R. Brinkman, B. Ellis, P. Haaland, D. Sarkar, J. Spidlen, E. Strain, R. Gentleman, Flowcore: a bioconductor package for high throughput flow cytometry, *BMC bioinformatics* 10 (1) (2009) 1–8. doi:10.1186/1471-2105-10-106.
20. C. E. Pedreira, E. S. Costa, S. Barrena, Q. Lecrevisse, J. Almeida, J. J. van Dongen, A. Orfao, Generation of flow cytometry data files with a potentially infinite number of dimensions, *Cytometry Part A: The Journal of the International Society for Analytical Cytology* 73 (9) (2008) 834–846. doi:10.1002/cyto.a.20608.
21. G. Lee, W. Finn, C. Scott, Statistical file matching of flow cytometry data, *Journal of biomedical informatics* 44 (4) (2011) 663–676. doi:10.1016/j.jbi.2011.03.004.
22. K. O'Neill, N. Aghaeepour, J. Parker, D. Hogge, A. Karsan, B. Dalal, R. R. Brinkman, Deep profiling of multitube flow cytometry data, *Bioinformatics* 31 (10) (2015) 1623–1631. doi:10.1093/bioinformatics/btv008.
23. A. Leite Pereira, O. Lambotte, R. Le Grand, A. Cosma, N. Tchitchek, Cytobackbone: an algorithm for merging of phenotypic information from different cytometric profiles, *Bioinformatics* 35 (20) (2019) 4187–4189. doi:10.1093/bioinformatics/btz212.
24. L. Balkay, *fca\_readfcs*, [Online; accessed 30-June-2020] (2020).
25. S. Wang, D. Li, X. Song, Y. Wei, H. Li. A feature selection method based on improved fisher's discriminant ratio for text sentiment classification. *Expert Systems with Applications*, 38(7) (2011) 8696-8702. doi:10.1016/j.eswa.2011.01.077.
26. M. Eveillard, V. Floc'h, N. Robillard, C. Debord, S. Wulleme, R. Garand, F. Rialland, C. Thomas, P. Peterlin, T. Guillaume, P. Moreau, P. Chevallier, M. C. Bene, CD38 expression in B-lineage acute lymphoblastic leukemia, a possible target for immunotherapy *Blood* 128 (22) (2016) 5268. doi:10.1182/blood.V128.22.5268.5268.
27. A. E. Bras, V. de Haas, A. van Stigt, M. Jongen-Lavrencic, H. B. Beverloo, J. G. Te Marvelde, C. M. Zwaan, J. J. van Dongen, J. H. Leusen, V. H. van der Velden, CD123 expression levels in 846 acute leukemia patients based on standardized immunophenotyping, *Cytometry Part B: Clinical Cytometry* 96 (2) (2019) 134–142. doi:10.1002/cyto.b.21745.
28. R. Shouval, J. A. Fein, B. Savani, M. Mohty, A. Nagler, Machine learning and artificial intelligence in haematology, *British Journal of Haematology* (2020) doi:10.1111/bjh.16915.
29. T. J. Keyes, P. Domizi, Y.-C. Lo, G. P. Nolan, K. L. Davis, A cancer biologist's primer on machine learning applications in high-dimensional cytometry, *Cytometry Part A* 97 (8) (2020) 782–799.



30. S. W. Krause, On its way to primetime: Artificial intelligence in flow cytometry diagnostics, *Cytometry Part A*. doi:10.1002/cyto.a.24191.
31. N. Aghaeepour, G. Finak, H. Hoos, T. R. Mosmann, R. Brinkman, R. Gottardo, R. H. Scheuermann, Critical assessment of automated flow cytometry data analysis techniques, *Nature Methods* 10 (3) (2013) 228–238. doi:10.1038/nmeth0513-445c.
32. Y. Saeys, S. V. Gassen, B. N. Lambrecht, Computational flow cytometry: helping to make sense of high-dimensional immunology data, *Nature Reviews Immunology* 16 (7) (2016) 449–462. doi:10.1038/nri.2016.56.
33. N. Radakovich, M. Nagy, A. Nazha, Machine learning in haematological malignancies, *The Lancet Haematology* 7 (7) (2020) e541–e550. doi:10.1016/S2352-3026(20)30121-6.
34. N. Mahmood, S. Shahid, T. Bakhshi, S. Riaz, H. Ghufuran, M. Yaqoob, Identification of significant risks in pediatric acute lymphoblastic leukemia (all) through machine learning (ml) approach, *Medical & Biological Engineering & Computing* (2020) 1–10. doi:10.1007/s11517-020-02245-2.
35. V. M. Pérez-García, S. Fitzpatrick, L. A. Pérez-Romasanta, M. Pesic, P. Schucht, E. Arana, P. Sánchez-Gómez, Applied mathematics and nonlinear sciences in the war on cancer, *Applied Mathematics and Nonlinear Sciences* 1 (2) (2016) 423–436. doi:10.21042/AMNS.2016.2.00036.
36. N. Radakovich, M. Nagy, A. Nazha, Artificial intelligence in hematology: Current challenges and opportunities, *networks* 2 (2020) 6. doi:10.1007/s11899-020-00575-4.
37. F. Malavasi, S. Deaglio, A. Funaro, E. Ferrero, A. L. Horenstein, E. Ortolan, T. Vaisitti, S. Aydin, Evolution and function of the adp ribosyl cyclase/CD38 gene family in physiology and pathology, *Physiological reviews* 88 (3) (2008) 841–886. doi:10.1152/physrev.00035.2007.
38. S. Deaglio, K. Mehta, F. Malavasi, Human CD38: A (r) evolutionary story of enzymes and receptors, *Leukemia research* 25 (1) (2001) 1–12. doi:10.1016/S0145-2126(00)00093-x.
39. S. Ibrahim, M. Keating, K.-A. Do, S. O'Brien, Y. O. Huh, I. Jilani, S. Lerner, H. M. Kantarjian, M. Albitar, CD38 expression as an important prognostic factor in b-cell chronic lymphocytic leukemia, *Blood, The Journal of the American Society of Hematology* 98 (1) (2001) 181–186. doi:10.1182/blood.v98.1.181.
40. A. Keyhani, Y. O. Huh, D. Jendiroba, L. Pagliaro, J. Cortez, S. Pierce, M. Pearlman, E. Estey, H. Kantarjian, E. J. Freireich, Increased CD38 expression is associated with favorable prognosis in adult acute leukemia, *Leukemia research* 24 (2) (2000) 153–159. doi:10.1016/S0145-2126(99)00147-2.
41. J. Naik, M. Themeli, R. de Jong-Korlaar, R. W. Ruiter, P. J. Poddighe, H. Yuan, J. D. de Bruijn, G. J. Ossenkoppele, S. Zweegman, L. Smit, T. Mutis, A.C.M. Martens N.W.C.J. van de Donk, R.W.J. Groen, CD38 as a therapeutic target for adult acute myeloid leukemia and t-cell acute lymphoblastic leukemia, *Haematologica* 104 (3) (2019) e100. doi:10.3324/haematol.2018.192757.
42. A. A. George, J. Franklin, K. Kerkof, A. J. Shah, M. Price, E. Tsark, D. Bockstoce, D. Yao, N. Hart, S. Carcich, R. Parkman, G. M. Crooks, K. Weinberg, Detection of leukemic cells in the CD34+ CD38- bone marrow progenitor population in children with acute lymphoblastic leukemia, *Blood, The Journal of the American Society of Hematology* 97 (12) (2001) 3925–3930. doi:10.1182/blood.V97.12.3925.
43. Y. Kong, S. Yoshida, Y. Saito, T. Doi, Y. Nagatoshi, M. Fukata, N. Saito, S. Yang, C. Iwamoto, J. Okamura, et al., CD34+ CD38+ CD19+ as well as CD34+ CD38- CD19+ cells are leukemia-initiating cells with self-renewal capacity in human b-precursor all, *Leukemia* 22 (6) (2008) 1207–1213. doi:10.1038/leu.2008.83.
44. D. C. Taussig, F. Miraki-Moud, F. Anjos-Afonso, D. J. Pearce, K. Allen, C. Ridler, D. Lillington, H. Oakervee, J. Cavenagh, S. G. Agrawal, T. A. Lister, J. G. Gribben, D. Bonnet, Anti-CD38 antibody-mediated clearance of human repopulating cells masks the heterogeneity of leukemia-initiating cells, *Blood*, 112 (3) (2008) 568–575. doi:10.1182/blood-2007-10-118331.
45. F. Lang, B. Wojcik, S. Bothur, C. Knecht, J. F. Falkenburg, T. Schroeder, H. Serve, O. G. Ottmann, M. A. Rieger, Plastic CD34 and CD38 expression in adult b-cell precursor acute lymphoblastic leukemia explains ambiguity of leukemia-initiating stem cell populations, *Leukemia* 31 (3) (2017) 731–734. doi:10.1038/leu.2016.315.
46. M. Ebinger, K.-E. Witte, J. Ahlers, I. Schäfer, M. André, G. Kerst, H.-G. Scheel-Walter, P. Lang, R. Handgretinger, High frequency of immature cells at diagnosis predicts high minimal residual disease level in childhood acute lymphoblastic leukemia, *Leukemia research* 34 (9) (2010) 1139–1142. doi:10.1016/j.leukres.2010.03.023.
47. T. V. Shman, L. V. Movchan, O. V. Aleinikova, Frequencies of immature CD34+ CD38- and CD34+ CD38- CD19+ blasts correlate with minimal residual disease level in pediatric b-cell precursor acute lymphoblastic leukemia, *Leukemia & lymphoma* 54 (11) (2013) 2560–2562. doi:10.3109/10428194.2013.778404.



48. J. Long, S. Liu, K. Li, X. Zhou, P. Zhang, L. Zou, High proportion of CD34<sup>+</sup>/CD38<sup>-</sup> cells is positively correlated with poor prognosis in newly diagnosed childhood acute lymphoblastic leukemia, *Leukemia & Lymphoma* 55 (3) (2014) 611–617. doi:10.3109/10428194.2013.807924.
49. Z. Jiang, D. Wu, S. Lin, P. Li, CD34 and CD38 are prognostic biomarkers for acute b lymphoblastic leukemia, *Biomarker research* 4 (1) (2016) 23. doi:10.1186/s40364-016-0080-5.
50. T. M. Owaidah, F. I. Rawas, N. B. Elkum. (2008). Expression of CD66c and CD25 in acute lymphoblastic leukemia as a predictor of the presence of BCR/ABL rearrangement. *Hematology/Oncology and Stem Cell Therapy*, 1(1), 34-37. doi:10.1016/S1658-3876(08)50058-6.
51. N. Guillaume, D. Penther, I. Vaida, B. Gruson, V. Harrivel, J.F. Claisse, J.C. Capiod, J.J. Lefrere, G. Damaj (2011). CD66c expression in B-cell acute lymphoblastic leukemia: strength and weakness. *International journal of laboratory hematology*, 33(1), 92-96. doi:10.1111/j.1751-553X.2010.01254.x.
52. N. Kiyokawa, K. Iijima, O. Tomita, M. Miharu, D. Hasegawa, K. Kobayashi, H. Okita, M. Kajiwarara, H. Shimada, T. Inukai, A. Makimoto, T. Fukushima, T. Nanmoku, K. Koh, A. Manabe, A. Kikuchi, K. Sugita, J. Fujimoto, Y. Hayashi, A. Ohara. (2014). Significance of CD66c expression in childhood acute lymphoblastic leukemia. *Leukemia Research*, 38(1), 42-48. doi:10.1016/j.leukres.2013.10.008.
53. E. Mejstříková, T. Kalina, J. Trka, J. Starý, O. Hrušák, Correlation of CD33 with poorer prognosis in childhood all implicates a potential of anti-CD33 frontline therapy, *Leukemia* 19 (6) (2005) 1092–1094. doi:10.1038/sj.leu.2403737.
54. U. Testa, E. Pelosi, A. Frankel, CD 123 is a membrane biomarker and a therapeutic target in hematologic malignancies, *Biomarker research* 2 (1) (2014) 1–11. doi:10.1186/2050-7771-2-4.
55. K. Liu, M. Zhu, Y. Huang, S. Wei, J. Xie, Y. Xiao, CD123 and its potential clinical application in leukemias, *Life sciences* 122 (2015) 59–64. doi:10.1016/j.lfs.2014.10.013.
56. M. Ruella, D. M. Barrett, S. S. Kenderian, O. Shestova, T. J. Hofmann, J. Perazzelli, M. Klichinsky, V. Aikawa, F. Nazimuddin, M. Kozlowski, J. Scholler, S.F. Lacey, J.J. Melenhorst, J.J.D. Morrisette, D.A. Christian, C. A. Hunter, M. Kalos, D.L. Porter, C. H. June, S.A. Grupp, S. Gill, Dual CD19 and CD123 targeting prevents antigen-loss relapses after CD19-directed immunotherapies, *The Journal of clinical investigation* 126 (10) (2016) 3814–3826. doi:10.1172/JCI87366.
57. E. Angelova, C. Audette, Y. Kovtun, N. Daver, S. A. Wang, S. Pierce, S. N. Konoplev, H. Khogeer, J. L. Jorgensen, M. Konopleva, P.A. Zweidler-McKay, L. J. Medeiros, H. M. Kantarjian, E. J. Jabbour, J. D. Khoury, CD123 expression patterns and selective targeting with a CD123-targeted antibody-drug conjugate (imgn632) in acute lymphoblastic leukemia, *haematologica* 104 (4) (2019) 749–755. doi:10.3324/haematol.2018.205252.
58. M. Djokic, E. Björklund, E. Blennow, J. Mazur, S. Söderhäll, A. Porwit, Overexpression of CD123 correlates with the hyperdiploid genotype in acute lymphoblastic leukemia, *haematologica* 94 (7) (2009) 1016–1019. doi:10.3324/haematol.2008.0002992.