Diagnosis of acute promyelocytic leukemia based on routine biological parameters using machine learning

by Estelle Cheli, Simon Chevalier, Olivier Kosmider, Marion Eveillard, Nicolas Chapuis, Adriana Plesa, Maël Heiblig, Lydie Andre, Jenny Pouget, Pascal Mossuz, Olivier Theisen, Vincent Alcazer, Dan Gugenheim, Nicolas Autexier, and Pierre Sujobert

Received: November 23, 2021.
Accepted: February 11, 2022.


Publisher's Disclaimer.
E-publishing ahead of print is increasingly important for the rapid dissemination of science. Haematologica is, therefore, E-publishing PDF files of an early version of manuscripts that have completed a regular peer review and have been accepted for publication. E-publishing of this PDF file has been approved by the authors. After having E-published Ahead of Print, manuscripts will then undergo technical and English editing, typesetting, proof correction and be presented for the authors' final approval; the final version of the manuscript will then appear in a regular issue of the journal. All legal disclaimers that apply to the journal also pertain to this production process.
Diagnosis of acute promyelocytic leukemia based on routine biological parameters using machine learning

Estelle Cheli¹, Simon Chevalier², Olivier Kosmider³, Marion Eveillard⁴, Nicolas Chapuis³, Adriana Plesa¹, Maël Heiblig⁵, Lydie Andre², Jenny Pouget⁴, Pascal Mossuz², Olivier Theisen³, Vincent Alcazer⁵, Dan Gugenheim⁶, Nicolas Autexier⁶, Pierre Sujobert¹,⁷

1. Hospices Civils de Lyon, Hôpital Lyon Sud, Service d'hématologie biologique.
2. CHU Grenoble Alpes, Service d’hématologie biologique
3. APHP, Hôpital Cochin, Service d’hématologie biologique
4. CHU Nantes, Service d’hématologie biologique
5. Hospices Civils de Lyon, Hôpital Lyon Sud, Service d'hématologie clinique.
6. Groupe onepoint, Bordeaux, France
7. Université Claude Bernard Lyon 1, Faculté de médecine et de maïeutique Lyon Sud Charles Mérieux, Lymphoma ImmunoBiology team.

Word count: 1595
2 figures
1 supplemental table and 2 supplemental figures
13 references

Short title: AI diagnosis of APL
Category: letter
Key words: Acute promyelocytic leukemia; artificial intelligence; machine learning
Disclosures: NA and DG were employed by Groupe onepoint when the work was conducted. The other authors have no conflict of interest to declare.
Fundings: this work was not funded by a specific grant.
Author Contributions:
PS, AP, EC and DG designed the study
EC, SC, OK, ME, NC, AP, PM, LA, JP, PM, OT and MH collected the data
NA, VA and DG developed the AI and the web interface

PS, EC, NA and DG analyzed the data

PS wrote the paper and all the authors made revisions and approved the final manuscript.

Acknowledgement:

The authors thanks Mrs Verena Landel from Hospices Civils de Lyon for English editing, as well as Dr Nicolas Duployez, Dr Damien Bodet, Dr Thomas Boyer and Dr Pascale Flandrin for their help in assessing AIPL performances on patients with t(8;16) AML.

Data sharing statement:

The data that support the findings of this study are available on request from the corresponding author.

Corresponding author:

Pierre Sujobert
Service d'hématologie biologique, Hôpital Lyon Sud, 165 chemin du grand Revoyet. 69310 Pierre Bénite, France
pierre.sujobert@chu-lyon.fr
Tel.: +33 4788615 / Fax: +33 478864104
Acute promyelocytic leukemia (APL) has become the most curable subtype of acute myeloid leukemia (AML) since the introduction of all-trans retinoic acid and arsenic trioxide.\textsuperscript{1} However, this disease is still characterized by a high rate of early death (10-17%), mainly due to severe coagulopathy.\textsuperscript{2,3} In order to avoid these early deaths, immediate treatment initiation is recommended, either with all-trans retinoic acid or with chemotherapy in case of hyperleucocytosis.\textsuperscript{4} Hence, a fast and accurate diagnosis is mandatory to allow early recognition and treatment of APL.

Cytology is the fastest technique for the diagnosis of APL, while the definitive confirmation requires the observation of the t(15;17) translocation or the PML-RARA fusion mRNA amplification, which induces further delay. Cytogenetic and molecular confirmation can be more difficult when other partners of RARA are implicated,\textsuperscript{5} and even more challenging in exceptional cases of viral insertion in the RARA gene, as recently described.\textsuperscript{6} The cytological diagnosis of APL is usually straightforward, when multiple bundles of Auer rods are observed in the blasts cells. However, the microgranular variant might be more difficult to diagnose, even for experienced hematopathologists. In some cases, myeloperoxidase deficiency in the blast cells further complicates the recognition of APL.\textsuperscript{7,8} Moreover, cytology requires a long training to recognize rare diseases such as APL, and this expertise is not always available.

We hypothesized that routine biological parameters might fuel an artificial intelligence to identify APL without a high level of cytological expertise. We collected 34 basic biological parameters in all the APL patients diagnosed in the Lyon University Hospital during the 2013-2020 period (n=76), and in patients with non-promyelocytic AML matched according to the year of diagnosis (n=146). Altogether, these patients constituted the cohort 1 (n=222). All the APL cases were confirmed by cytogenetic and/or RT-qPCR amplification of the PML-RARA fusion transcript. The biological parameters were measured during the first two days of hospital referral, before any treatment initiation. Missing data were imputed by the
variable’s median value. The basic hematology and hemostasis parameters were available for most of the patients, but there were more missing data concerning the biochemical parameters (supplemental table 1). The cohort was randomly split into a training (80%, n=177) and a test cohort (20%, n=45). Different classification algorithms were then compared (XGBoost, random forest, gradient boosting classifier, adaboost classifier, decision tree, logistic regression and support vector machine), considering APL diagnosis as a binary outcome and using 5X cross-validation to select the more stable models. No normalization of the data was used, because both strategies tested (StandardScaler, MinMaxScaler) had a negative impact on the performances of the algorithm. Hyperparameter tuning was performed using GridSearchCV. All analyses were performed using Python v3.7. Among the different artificial intelligence strategies tested, XGBoost’s gradient boosting algorithm achieved the highest performances in the test cohort, with an area under the receiver operator curve (ROC) of 0.95 (figure 1A, see supplemental material for methodological details). Of note, learning curves reached a plateau with 80-100 patients, meaning that no major refinements is expected with an increase in the size of the cohort (data not shown). Using this model, we established Artificial Intelligence for Promyelocytic Leukemia (AIPL), an open-source tool with a graphical user interface to evaluate the probability of APL diagnosis (https://github.com/Nico-Facto/Leukemia-Apl-Classification) and propose a ready-to-use web application (https://share.streamlit.io/nico-facto/leukemia-apl-classification/main/Leucemie_app.py). The 8 parameters required to run AIPL are the following: age, white blood cells (absolute value), lymphocyte (% of total leucocytes), neutrophil polynuclear count (absolute value), mean corpuscular volume (MCV), mean corpuscular hemoglobin concentration (MCHC), prothrombin time ratio, and fibrinogen concentration.
To validate the AIPL tool, its performances were assessed in 3 independent retrospective validation cohorts from 3 other hospitals (cohorts 2, 3, and 4) which comprised 44 (including 15 APL), 258 (including 46 APL), and 63 (including 32 APL) patients, respectively. A prospective cohort (cohort 5) was also collected in the Lyon University Hospital with 50 (including 10 APL) new AML diagnoses referred during a 6-month period. AIPL showed a very high discrimination ability both in the merged (n= 415 patients, AUC =0.96, figure 1B) and in individual cohorts (figure 1C). Importantly, AIPL output is not only a classification (APL vs non-APL), but also a confidence score reflecting how much the conclusion can be trusted. As expected, the confidence score was significantly higher in cases where the prediction was correct compared to cases where the AIPL prediction failed (mean 95% vs 85%, Mann-Whitney test p<0.0001, figure 2A). Hence, the AIPL confidence score could be used to determine for which patient the prediction of AIPL is reliable in routine use. For 244 (59%) patients with a high confidence score (above 99%), the accuracy of AIPL was 99.5% (only one false negative case, i.e. an APL patient wrongly classified as non-APL AML). For 114 (27%) patients with an intermediate confidence score (between 85 and 99%), the accuracy was 85% (7 false negative and 10 false positive cases, i.e. a non-APL AML patient wrongly classified as APL), and for 57 (14%) patients with a low confidence score (below 85%), the accuracy dropped to 68% (8 false negative and 10 false positive cases) (figure 2B).

As the data from the retrospective cohorts were acquired on different analyzers, it was possible to assess the impact of the variability in analytical techniques on AIPL performances. When comparing different time periods determined by changes in analyzers, there was no significant variation in the level of confidence scores of AIPL, suggesting that this approach is robust to variations due to analytical processes (supplemental figure 1).

Of note, 16 cases of the microgranular variant of APL were identified in the validation cohorts. Their AIPL confidence scores tended to be lower than the confidence scores of
classical APL (86 vs 92%, ns, supplemental figure 2), suggesting that further algorithm training using microgranular variants might be interesting. Importantly, AIPL correctly identified the six microgranular cases with a confidence score above 99%. We also assessed the performances of AIPL in patients with other differential diagnosis of APL: aplastic anemia (n=10), acute lymphoblastic leukemia (n=28) and AML with t(8;16), a rare subtype of AML with clinical presentation often resembling APL (n=9). AIPL classified only one case aplastic anemia and one case of acute lymphoblastic leukemia as APL, but with confidence score below 87%. Hence, with the proposed threshold of confidence score of 99%, there was no false positive diagnostic of APL in these challenging differential diagnoses.

To further interpret the predictions from AIPL, Shapley additive explanation (SHAP) was used to illustrate the impact of the different parameters according to their value obtained from the individual cases of all the validation cohorts. In figure 2C, the parameters are ordered from top to bottom according to their importance in the classification, and each individual measure is colored according to its impact on the final classification. The high performances of AIPL rely on some expected parameters such as fibrinogen, prothrombin time ratio, or polynuclear neutrophil count (figure 2C). Unexpectedly, two parameters of red blood cells (MCV and MCHC) were highly discriminant between APL and non-APL AML, even if the mean values of these parameters remained in the normal ranges (mean value of MCV 89 vs 96 fl, mean value of MCHC 349 vs 334 g/L in APL and non-APL cases, respectively). This observation, together with a report of PML-RARA expression in burst forming unit-erythroid (BFU-E) derived from APL patients, raises the hypothesis that PML-RARA cells contribute to erythropoiesis.

To conclude, this work demonstrates that machine learning based on routine biological parameters provides a fast and accurate help for the diagnosis of APL in the majority of cases.
Given the unmet need to improve the reliability of APL diagnosis, other strategies based on incorporation of cell population data generated during complete blood cell count with cytometry based analyzers,\textsuperscript{12} or deep learning analysis of blood smears, have also been developed.\textsuperscript{13} As these approaches rely on parameters not used in AIPL, combining them with AIPL could help further increase diagnostic accuracy. Another interesting possibility could be the addition of other biological parameters such as the fibrinolysis marker D-Dimers, which were not included in this retrospective study due to excessive missing data.

AIPL might represent a very important complement to cytological expertise, and could allow early diagnosis of APL in settings where this expertise is not available on a 24/7 basis, or not available at all, such as in developing countries. The consequences of misclassification could be excessive treatment with ATRA in case of false positive result, or delay in ATRA initiation in case of false negative result. Using AIPL with the proposed threshold of 99% of confidence score, this risk is very low but should not be forgotten. All-trans retinoic acid could hence be initiated in patients with a high probability of APL according to AIPL prediction without waiting for diagnostic confirmation in specialized laboratories, thus preventing early death from coagulopathy. However, an important limitation of AIPL is that its ability to distinguish APL from other differential diagnoses such as acute lymphoblastic leukemia or aplastic anemia has not been formally assessed in this study. To make this tool available, a web user interface has been created (available at https://share.streamlit.io/nico-facto/leukemia-apl-classification/main/Leucemie_app.py) to use AIPL in the case of patients with myeloid blast on the blood smear. It allows to instantaneously classify patients as APL or non-APL and provides a confidence score. Of course, this result does not supplant the need to evaluate the bone marrow and to formally demonstrate the presence of the t(15;17) translocation or the PML-RARA fusion transcript. Given the importance of early treatment initiation in these patients, we hope that AIPL will contribute to decrease early mortality in APL patients.
References


Figure legends:

Figure 1: Development of AIPL for the diagnosis of acute promyelocytic leukemia
A: Receiver-operator characteristics (ROC) curve of the XGBoost method in the test cohort.
B: Receiver-operator characteristics (ROC) curve in all the patients from the validation cohorts
C: Area under the curve of the ROC curves in each validation cohort, separately.

Figure 2: Validation of AIPL
A: Mean confidence score of AIPL in cases correctly or wrongly classified in all the validation cohorts.
B: AIPL performances according to the confidence score.
C: Shapley Additive Explanation (SHAP) representation of the parameters used by AIPL. The parameters are ordered from top to bottom according to their importance in the classification, and each individual measure is colored according to its impact on the final classification (the more a point is red, the more it favors the classification as APL)
Figure 1

A

Roc curve cv

AUC = 0.95

B

Roc curve cv

AUC = 0.96

C

cohort 2

n=44

cohort 3

n=258

cohort 4

n=63

cohort 5

n=50
### Supplemental table 1: Missing data in the training cohort

<table>
<thead>
<tr>
<th>Features</th>
<th>Missing data, n (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Demographic characteristics</strong></td>
<td></td>
</tr>
<tr>
<td>Sex</td>
<td>0 (0)</td>
</tr>
<tr>
<td>Age at diagnosis</td>
<td>0 (0)</td>
</tr>
<tr>
<td><strong>Hematology</strong></td>
<td></td>
</tr>
<tr>
<td>Leucocytes (x10⁹/L)</td>
<td>1 (0.5)</td>
</tr>
<tr>
<td>Red Blood Cell (x10¹²/L)</td>
<td>6 (2.7)</td>
</tr>
<tr>
<td>Hemoglobin (g/L)</td>
<td>1 (0.5)</td>
</tr>
<tr>
<td>Hematocrit (%)</td>
<td>5 (2.25)</td>
</tr>
<tr>
<td>Mean Corpuscular Volume (fL)</td>
<td>4 (1.8)</td>
</tr>
<tr>
<td>Mean Corpuscular Hemoglobin (pg)</td>
<td>3 (1.3)</td>
</tr>
<tr>
<td>Mean Corpuscular Hemoglobin Concentration (g/L)</td>
<td>5 (2.25)</td>
</tr>
<tr>
<td>Red blood cell Distribution Width-CV (%)</td>
<td>6 (2.7)</td>
</tr>
<tr>
<td>Neutrophils (x10⁹/L)</td>
<td>7 (3.15)</td>
</tr>
<tr>
<td>Eosinophils (x10⁹/L)</td>
<td>9 (4.05)</td>
</tr>
<tr>
<td>Basophils (x10⁹/L)</td>
<td>9 (4.05)</td>
</tr>
<tr>
<td>Lymphocytes (x10⁹/L)</td>
<td>9 (4.05)</td>
</tr>
<tr>
<td>Monocytes (x10⁹/L)</td>
<td>9 (4.05)</td>
</tr>
<tr>
<td>Blasts (%)</td>
<td>12 (5.4)</td>
</tr>
<tr>
<td>Platelets (x10⁹/L)</td>
<td>2 (0.9)</td>
</tr>
<tr>
<td>Mean Platelet Volume (fL)</td>
<td>35 (15.7)</td>
</tr>
<tr>
<td>Reticulocytes (x10⁹/L)</td>
<td>108 (48.6)</td>
</tr>
<tr>
<td><strong>Hemostasis</strong></td>
<td></td>
</tr>
<tr>
<td>Prothrombin ratio (%)</td>
<td>3 (1.3)</td>
</tr>
<tr>
<td>Activated Partial Thromboplastin Time</td>
<td>2 (0.9)</td>
</tr>
<tr>
<td>Fibrinogen (g/L)</td>
<td>8 (3.6)</td>
</tr>
<tr>
<td>Fibrin Degradation Products (µg/mL)</td>
<td>194 (87.4)</td>
</tr>
<tr>
<td><strong>Biochemistry</strong></td>
<td></td>
</tr>
<tr>
<td>LDH (U/L)</td>
<td>30 (13.5)</td>
</tr>
<tr>
<td>Calcium (mmol/L)</td>
<td>4 (1.8)</td>
</tr>
<tr>
<td>Phosphorus (mmol/L)</td>
<td>23 (10.3)</td>
</tr>
<tr>
<td>Ferritin (µg/L)</td>
<td>83 (37.4)</td>
</tr>
<tr>
<td>C Reactive Protein (mg/L)</td>
<td>21 (9.5)</td>
</tr>
<tr>
<td>Urea (mmol/L)</td>
<td>3 (1.3)</td>
</tr>
<tr>
<td>Creatinine (µmol/L)</td>
<td>2 (0.9)</td>
</tr>
<tr>
<td>Uric Acid (µmol/L)</td>
<td>22 (9.9)</td>
</tr>
</tbody>
</table>
Supplemental figure 1: Value of the confidence score for each patient according to the time of diagnosis in the validation cohort n°3. The analyzers automats used for hematology, hemostasis, and biochemistry were the following: period 1: Siemens Dimension Vista, StaR, Beckman Coulter LH780; period 2: Siemens Dimension Vista, StaRevolution, Sysmex XE5000 and XE2100 D; period 3: Siemens Atellica, StaRMax3 and Sysmex XN. The differences are non-significant (Kruskal-Wallis test).

Supplemental figure 2: Variations in AIPL confidence scores according to the cytological subtype of APL.