

ARTICLE OPEN



CHRONIC MYELOID LEUKEMIA

Evaluating a predictive model of tyrosine kinase inhibitor therapy failure in a European-type cohort: a step towards population-specific tools

Sylvain Moinard ^{1,2,20}, Benjamin Lebecque ^{1,2,3,20}, Tom Lachaise ^{1,2}, Hyacinthe Johnson-Ansah^{3,4}, Charlotte Doublet^{3,5}, Gabrielle Roth-Guepin ^{3,6}, Françoise Rigal-Huguet^{3,7}, Lydia Roy^{3,8}, Anne Parry⁹, Mathieu Meunier ^{3,10}, Amélie Penot ^{3,11}, Laurence Legros ^{3,12}, Vanessa Pante¹³, Martine Escoffre-Barbe^{3,14}, Philippe Rousselot^{3,15}, Guillaume Denis ¹⁶, Hélène Monjanel ¹⁷, Dalil Hamroun ¹⁸, Abdessattar Khlaifia^{1,2}, Joévin Besombes ^{1,2}, Bruno Pereira¹⁹, Céline Bourgne ^{1,2} and Marc G. Berger^{1,2,3}✉

© The Author(s) 2025

Predicting therapeutic failure in patients with chronic phase-chronic myeloid leukemia (CP-CML) treated with tyrosine kinase inhibitors (TKI) remains a major challenge for personalized care management. The Sokal and EUTOS long-term survival scores were designed to predict CML-related mortality, but are also used to guide therapeutic choices, despite their poor performance for this purpose. A recent study proposed a refined predictive model of therapy failure specifically tailored for patients treated with imatinib and second-generation TKIs that showed promising results in a Chinese cohort. The present study evaluated the performance and applicability of this predictive model in a real-world, multicenter cohort from the French CML Observatory. The key differences identified between the Chinese and French cohorts (age, baseline hemoglobin levels, and treatment regimens) likely influenced the model performance. Specifically, the new model did not allow for discriminating risk groups effectively in the French cohort. However, the model reconstruction using this cohort identified other predictive variables (sex, leukocytosis, comorbidities, high-risk additional chromosomal abnormalities) that better stratified patients at risk of therapy failure. Our findings highlight the influence of demographic and clinical differences on predictive models and emphasize the need for local or population-specific tools to optimize risk stratification and therapeutic decision-making in CP-CML.

Leukemia (2025) 39:2375–2383; <https://doi.org/10.1038/s41375-025-02703-6>

INTRODUCTION

Tyrosine kinase inhibitors (TKI) are an unprecedented breakthrough for chronic phase-chronic myeloid leukemia (CP-CML) management, but the proportion of patients with insufficient therapeutic response, according to the European LeukemiaNet (ELN) recommendations [1], is still high. The early identification of at-risk patients is critical for therapeutic decision-making optimization.

Prognostic scores originally designed to predict CML-related mortality risk, such as the Sokal score [2] in the pre-TKI era and the EUTOS long-term survival (ELTS) score [3] in the TKI era, are also used to predict resistance to therapy [4, 5]. However, their performance in predicting the optimal response is low [6].

Recently, Zhang et al. developed a refined predictive model of TKI failure using classical demographic and clinical variables at diagnosis. This model, initially developed using data of imatinib-treated patients [7], was then adapted to a cohort of patients treated with second-generation TKIs (2G-TKIs) and showed a promising performance [8]. Nevertheless, this score was established in an ethnically homogeneous Chinese cohort. Therefore, as stressed by the authors in their conclusion and recently by Lauseker et al. [9], its performance needs to be evaluated in non-Asian cohorts. Here, we assessed the performance and applicability of this predictive model in a real-world French cohort of patients with CP-CML. We then reconstructed the model using our data and the

¹service d'Hématologie Biologique, CHU Clermont-Ferrand, Clermont-Ferrand, France. ²EA7453 CHELTER, Université Clermont Auvergne, Clermont-Ferrand, France. ³France Intergrupee de la leucémie myéloïde chronique Fi-LMC, Lyon, France. ⁴service d'Hématologie, CHU de Basse Normandie, Caen, France. ⁵service d'Hématologie, CHU Saint-Etienne, Saint-Etienne, France. ⁶service d'Hématologie, CHRU Brabois, Vandœuvre-lès-Nancy, France. ⁷service d'Hématologie, Institut Universitaire du Cancer Toulouse Oncopôle, CHU Toulouse, Toulouse, France. ⁸service d'Hématologie, APHP Henri Mondor, Créteil, France. ⁹service d'Hématologie, CH Annecy-Genevois, Pringy, France. ¹⁰service d'Hématologie, CHU Grenoble Alpes, Grenoble, France. ¹¹service Hématologie clinique et thérapie cellulaire, CHU Limoges, Limoges, France. ¹²service Hématologie clinique ambulatoire, APHP Hôpital Bicêtre, Kremlin-Bicêtre, France. ¹³service Oncologie - chimiothérapie, CH Emile Roux, Le Puy-en-Velay, France. ¹⁴Hôpital de Pontchaillou, service d'Hématologie Clinique, CHU de Rennes, Rennes, France. ¹⁵service d'Hématologie-Oncologie, CH Versailles, Versailles, France. ¹⁶Médecine Interne et Hématologie, CH Rochefort, Rochefort, France. ¹⁷Médecine Interne, endocrinologie, diabétologie et neurologie, CH Henri Mondor, Aurillac, France. ¹⁸Direction de La Recherche et de L'Innovation, CHRU de Montpellier, Montpellier, France. ¹⁹secteur de Biométrie et Médico-Economie, DRCI, CHU Clermont-Ferrand, Clermont-Ferrand, France. ²⁰These authors contributed equally: Sylvain Moinard, Benjamin Lebecque. ✉email: mberger@chu-clermontferrand.fr

Received: 28 April 2025 Revised: 11 June 2025 Accepted: 7 July 2025

Published online: 1 August 2025

same procedure as Zhang et al. to identify key variables that could improve the precision of therapy failure prediction.

MATERIALS/SUBJECTS AND METHODS

Data source and patient enrollment

We selected patients who received a diagnosis of CP-CML from the French CML Observatory (NCT05963061), a multicenter database launched in 2013 to collect real-world data (Supplementary Information). This secure platform is hosted by MIPiH and authorized by the French data protection authority (CNIL, no. 914456). Patients are included after informed consent, and pseudonymized data are collected retrospectively and prospectively. The CML Observatory is promoted by Clermont-Ferrand University Hospital. As standardized quantification of the BCR::ABL transcript only became available in 2010, we included patients with a diagnosis obtained between 2010 and 2024. A potential center effect, resulting from the discretion granted to investigators in selecting patient records, was mitigated by the participation of 16 centers; this effect was further addressed in the statistical analysis.

Like in the study by Zhang et al. [8], we excluded patients with more than 6 months between diagnosis and TKI initiation, advanced-phase-CML diagnosis, missing key covariates, irregular response monitoring, lost to follow-up, or with non-e14a2 and/or e13a2 transcripts. For the present study, we closed the database on January 8, 2025.

As in Zhang et al., we defined the molecular responses (MMR, MR4, MR4.5), complete cytogenetic response, and complete hematological response according to the ELN recommendations [1], and TKI failures according to the “failure” milestones in the 2020 ELN recommendations (Supplementary Information). We calculated failure-free survival from the date of TKI therapy start to the date of the first therapy failure or censored patients at the transplant, death, or last follow-up date. We excluded patients with several observed TKI failures from further analysis after the first therapy failure.

Experimental strategy

To evaluate the performance of the score by Zhang et al. in this European-type population, we defined a three-step strategy: (1) External validation: We first applied the original predictive model developed by Zhang et al. [8]. (Model 1) to our cohort from the French CML Observatory. Given the differences in patient characteristics, particularly regarding ethnicity, age, and genetic factors, between our cohort and the original Chinese cohort, we anticipated potential discrepancies in the model performance. We hypothesized that the influence of individual variables might differ, and added two new steps: (2) Score recalibration: Using the same six key variables identified by Zhang et al. [8], we recalculated the weights of the Fine-Gray model [10] based on our French cohort to create a redefined model (Model 2); and (3) Model optimization: We explored the twelve candidate variables initially considered by Zhang et al. and selected the most relevant ones for predicting therapy failure to build an optimized Fine-Gray model (Model 3).

Statistical analyses

The analysis pipeline follows the one by Zhang et al. We used descriptive statistics to summarize baseline covariates (Supplementary Information). We calculated the cumulative incidences of therapy failure using the Fine-Gray model and compared them with the Gray test [11]. Due to the low number of patients who died ($n = 31$; 5%) and the fact that therapy failures and deaths occurred at different timescales (25% of deaths and therapy failures, quartile 1, occurred within 26 months and within 3 months, respectively; 75% of deaths and therapy failures, quartile 3, occurred within 84 months and 27 months,

respectively), we did not treat deaths as a competing risk in the main analysis. We also performed a sensitivity analysis, taking into account death as a competing risk (Supplementary Information).

After applying the predictive score developed by Zhang et al. (Model 1), we developed Model 2 by replicating Zhang et al.’s methodology using the same six covariates [sex, age, hemoglobin level, percentage of blood blasts, spleen size below the costal margin, and presence of high-risk additional chromosomal abnormalities (ACAs)] to build a new Fine-Gray model based on data of our French cohort. Then, we built Model 3 using the twelve candidate variables initially considered by Zhang et al.: the six previous ones, and also white blood cell count, platelet count, basophils, eosinophils, presence of comorbidities, and first-line TKI. We selected four relevant variables (sex and high-risk ACAs, like in the original study, as well as white blood cell count and presence of comorbidities) for the global Fine-Gray model based on the Akaike Information Criterion (AIC) to build a robust predictive model. We carried out sensitivity analyses, taking the competing risk (death) into account or not, which influences the selection of comorbidities in the choice of final model. We chose to retain them in Model 3 due to their clinical relevance. For Models 2 and 3, we chose the cutoffs to determine risk groups by bootstrapping (Supplementary Information).

We assessed the predictive performance of the three models by comparing the hazard ratios between the risk groups and on the basis of the receiver-operating characteristic (ROC) curves after 1, 3, and 5 years of TKI therapy. We compared the areas under the ROC curves (AUC) using the DeLong method [12].

We used propensity score matching with replacement to compare failure-free survival in the first-line imatinib and first-line 2G-TKI groups (Supplementary Information).

We used R version 4.4.1 (R Core Team, Vienna, Austria) for the statistical analyses and graph generation.

RESULTS

Characteristics of the French cohort

Using the inclusion/exclusion criteria by Zhang et al. [8] and after excluding patients with a diagnosis before 2010 (due to the lack of inter-laboratory standardization of BCR::ABL1 transcript quantification before that date) and also patients with missing data and those who received non-TKI treatments, we identified 601 patients from 16 centers in the CML Observatory (Fig. 1) treated with: imatinib ($n = 397$), nilotinib ($n = 143$), bosutinib ($n = 48$), and dasatinib ($n = 13$). Among these 601 patients, 188 (31.3%) experienced treatment failure.

The main differences between our cohort and the Chinese training cohort were (Table 1): older median age (58 [47, 70] vs. 40 [30–52] years), lower percentage of men (54% vs. 61%), more patients with intermediate-risk Sokal (46% vs. 28%) and high-risk ELTS (14% vs. 9%) scores, lower initial white blood cell count (64 [30, 139] vs. 122 [47, 235] $\times 10^9/L$) and higher hemoglobin concentration (126 [111, 139] vs. 115 [97, 132] g/L). More patients received 2G-TKIs in our cohort (34% vs. 21%). Bosutinib-treated patients were present only in our cohort (8%), but no patient received flumatinib (vs. 3%). Moreover, in our cohort, 30% of patients had comorbidities (vs. 36%), and 5% had high-risk ACAs (vs. 2%). Treatment failure was reported for 188 patients (31% vs. 25%).

Application of the score formula developed by Zhang et al. (Model 1)

First, we applied the predictive formula validated by Zhang et al. (Model 1). This classified patients into low ($n = 204/601$, 33.9%), intermediate ($n = 346/601$, 57.6%), and high risk (51/601, 8.5%) (Supplementary Tables 1 and 2). Unlike in the original Chinese cohort, the predictive model could efficiently discriminate only the low- and high-risk groups (Fig. 2A) in our overall cohort

($p = 0.003$), as well as in the first-line imatinib and 2G-TKI groups (Fig. 2B, C) (Supplementary Tables 1 and 2). Despite this significant result, the hazard ratios obtained using the low-risk group as reference were low compared with those reported by Zhang et al.: 1.2 (95% CI: 0.9–1.7, $p = 0.2$) vs. 5.4 (95% CI: 4.1–7.2, $p < 0.001$) for the intermediate-risk group and 2.0 (95% CI: 1.2–3.2, $p = 0.006$) vs. 12.1 (95% CI: 8.8–16.6, $p < 0.001$) for the high-risk group. This suggests a low predictive performance of these risk groups in our French cohort (Supplementary Fig. 1).

Recalibration of the score using the French cohort and the same six variables (Model 2)

Given the divergence in predictive performance between the French and Chinese cohorts, we recalculated the weights in the score formula using the same six baseline variables as Zhang et al. but using data from our French cohort (Model 2) (Fig. 2D–F, Table 2). This classified patients into low ($n = 234/601$, 39%), intermediate ($n = 142/601$; 24%) and high risk (225/601; 37%) (Supplementary Tables 1 and 2, Supplementary Fig. 2). Model 2 provided a more consistent stratification, with significant discrimination between the low- and intermediate-risk groups ($p = 0.01$) and the low- and high-risk groups ($p < 10^{-5}$), regardless of the first-line TKI. However, the overall predictive performance remained inferior to that reported for the original Chinese cohort. Indeed, the hazard ratios for the intermediate-risk and high-risk groups remained lower than those reported by Zhang et al.: 1.5 (95% CI: 1.0–2.2, $p = 0.039$) vs. 5.4 for the intermediate-risk group and 2.1 (95% CI: 1.5–3.0, $p < 0.001$) vs. 12.1 for the high-risk group (Supplementary Fig. 1).

Using the twelve initial variables to recalculate the predictive score (Model 3)

These results prompted us to analyze the 12 covariates initially used by Zhang et al., based on the hypothesis that some covariates might carry a different weight in our cohort (Model 3) (Table 2). The results of the univariate analyses are provided in Supplementary Table 3. Using a stepwise variable selection for the Fine-Gray regression model, we identified four variables independently associated with the risk of therapeutic failure: sex and high-risk ACAs (like in the original study), as well as leukocytosis and presence of comorbidities, two additional factors not retained in the initial model by Zhang et al.

Model 3 produced a risk stratification similar to Model 2: 54% of patients were classified as low risk, 24% as intermediate risk, and 21% as high risk (Supplementary Tables 1 and 2, Supplementary Fig. 3). The model showed significant discrimination between the low- and intermediate-risk groups ($p = 0.01$) and between the low- and high-risk groups ($p < 10^{-5}$) (Fig. 2G). However, the hazard ratios for the intermediate-risk and high-risk groups remained lower than those determined by Zhang et al.: 1.7 (95% CI: 1.2–2.5, $p = 0.003$) vs. 5.4 for the intermediate-risk group and 2.1 (95% CI: 1.5–3.0, $p < 0.001$) vs. 12.1 for the high-risk group (Supplementary Fig. 1).

Moreover, after propensity score matching, the risk group discrimination differed between patients treated with imatinib and with 2G-TKIs due to the lower therapy failure rate in the 2G-TKI group (Supplementary Fig. 4 and Supplementary Table 4). The

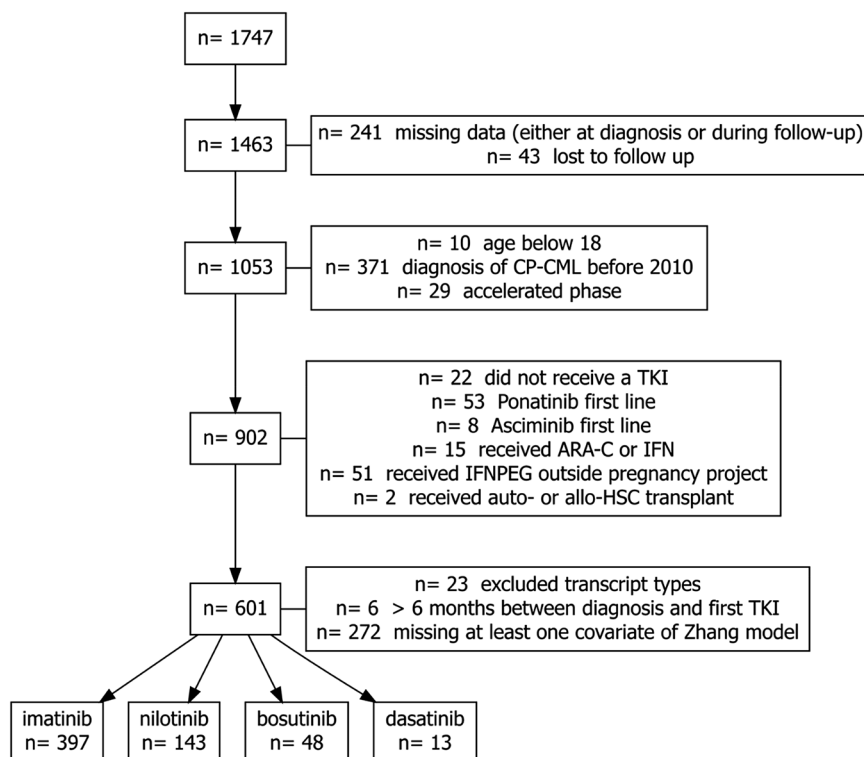


Fig. 1 Study flowchart of patient selection from the French CML Observatory database. Among the 1747 patients initially registered, 1463 were retained after excluding those with missing data (either at diagnosis or during follow-up, $n = 241$) or lost to follow-up ($n = 43$). Subsequent exclusions included patients younger than 18 years ($n = 10$), with a diagnosis of CP-CML before 2010 ($n = 371$), or in the accelerated phase at diagnosis ($n = 29$), resulting in 1053 patients. Moreover, 151 patients were excluded because they received non-standard treatments: no TKI treatment ($n = 22$), first-line ponatinib ($n = 53$), asciminib ($n = 8$), ARA-C or interferon (IFN) ($n = 15$), PEG-IFN outside a pregnancy project ($n = 51$), transplant ($n = 2$). Lastly, 23 patients were excluded due to transcript type, 6 due to treatment initiation >6 months after diagnosis, and 272 because they missed covariates analyzed in Zhang et al. The analyzed cohort included 601 patients treated with first-line imatinib ($n = 397$), nilotinib ($n = 143$), bosutinib ($n = 48$), or dasatinib ($n = 13$).

Table 1. Patients' characteristics in the French multicenter cohort and the training cohort in the study by Zhang et al.

	French cohort (N = 601)	Chinese training cohort (N = 1955)
Age, years, median (IQR)	58 (47–70)	40 (30–52)
Male sex, n (%)	323 (54)	1195 (61)
Spleen size, cm below costal margin, median (IQR)	0 (0–3)	3 (0–10)
White blood cells, ×10E + 9/L, median (IQR)	64 (30–139)	122 (47–235)
Hemoglobin, ×10E + 9/L, median (IQR)	126 (111–139)	115 (97–132)
Platelets, ×10E + 9/L, median (IQR)	389 (248–600)	410 (270–635)
Blood blasts, %, median (IQR)	0 (0–1.7)	1 (0–3)
Blood basophils, %, median (IQR)	3 (2–5)	5 (2–8)
Blood eosinophils, %, median (IQR)	2 (1–3)	2 (1–4)
Sokal risk, n (%)		
Low	206 (34)	819 (42)
Intermediate	278 (46)	555 (28)
High	113 (19)	394 (20)
Unknown	4 (0.7)	187 (10)
ELTS risk, n (%)		
Low	321 (53)	1115 (57)
Intermediate	192 (32)	471 (24)
High	83 (14)	182 (9)
Unknown	5 (0.8)	187 (10)
Ph+ ACAs, n (%)	70 (12)	68 (3)
High-risk ACAs, n (%)	32 (5)	41 (2)
Comorbidity(ies), n (%)	180 (30)	700 (36)
Initial TKI, n (%)		
Imatinib	397 (66)	1539 (79)
2 G TKI	204 (34)	416 (21)
Nilotinib	143 (24)	280 (14)
Bosutinib	48 (8)	0 (0)
Dasatinib	13 (2)	72 (4)
Flumatinib	0 (0)	64 (3)
Follow-up, months, median (IQR)	47 (19–89)	56 (30–91)

predictive performance of the Models remained similar after propensity score matching.

Comparison of the performance of the three models and the Sokal and ELTS prognostic scores

We used alluvial plots (Fig. 3) to illustrate the patient flow between the Sokal, ELTS, and Models 1, 2, and 3 risk groups. Model 1 gave results relatively close to those of the Sokal and ELTS scores; for example, the high-risk group in Model 1 was small (8%) and largely corresponded to patients classified as high risk by the traditional scores. For Models 2 and 3, the relationships were more complex, and notable proportions of patients shifted between groups in both directions. Some patients were classified as high risk by the models developed in this study, but as low risk by the

traditional scores (up to 60 patients between the Model 2 high-risk group and the ELTS low-risk group, Fig. 3). Model 2 identified the largest high-risk group (37%), and Model 3 defined an intermediate-size high-risk group (21%) that better matched the percentage of patients with therapy failure (31%) in our cohort, unlike the high-risk score by Zhang et al. and the ELTS score (8% and 14%, respectively) [3, 8] (Fig. 3).

To compare the predictive performance of the ELTS and Models 1, 2, and 3 scores, we calculated their AUC values, which were all <0.68 (Fig. 4A–C), regardless of the first-line TKI (Fig. 4D–I). Model 3 tended to be slightly more effective, but the difference was not significant. Results were similar to the ELTS score performance (Figs. 2J–L and 4), including the relationship with the cumulative incidence of molecular responses (Supplementary Fig. 5), confirming previously published results in this cohort [6].

DISCUSSION

Developing predictive scores for the therapeutic response or failure based on parameters collected during the early phase of the disease remains challenging, although they would help to predict TKI efficacy and to personalize the management of patients with CP-CML. In clinical practice, the Sokal score and, more recently, the ELTS score are often taken into account when choosing the first-line TKI. However, these scores were developed to assess the CML-related risk of mortality, and their performance in predicting the response to TKIs is poor, as we demonstrated using data from the French CML Observatory [6]. Consequently, the approach taken by Zhang et al. to define a score more suitable to predict the therapeutic response is very relevant. In the Chinese cohort, the new score defined by this group appears particularly effective. However, in our study, this score did not give the same results and did not clearly distinguish between risk groups, as indicated by the modest hazard ratios compared with those obtained by Zhang et al. This underlines the need to verify the performance of a score in different populations, as suggested by Zhang et al. [8] and recently recommended [9].

This difference may be partly explained by the two cohorts' characteristics. For example, Asian patients treated with TKIs harbor a higher proportion of BCR::ABL mutations associated with resistance to imatinib [13]. Moreover, the lower hemoglobin concentration in the Chinese cohort is in line with the lower reference values in Asian populations [14], but could also be associated with the likelihood of achieving optimal responses [15]. This observation could be related to underlying conditions, such as thalassemia, the prevalence of which is approximately 50 times higher in Southeast Asian populations. However, Zhang et al. did not provide information on the influence of a thalassemia trait on TKI tolerance. Another difference was the median age at disease onset, likely due to unknown genetic characteristics. Lastly, potential pharmacokinetic differences cannot be ruled out because the same recommended dose has been used for Western and Asian patients despite their lower individual body weight [13]. This observation highlights the need to assess this type of score in patient populations with different ethnic backgrounds, as emphasized by Zhang and coworkers in their conclusion.

Therefore, we hypothesized that the various, fairly standard variables used to calculate the score might influence the risk of therapy failure differently in different populations, and we re-evaluated the score using the six variables selected by Zhang et al. (Model 2). However, the performance of this score remained poor in our cohort. The relative importance of variables differed between Model 1 (mainly age and high-risk ACAs, all variables were significantly associated with the therapy failure risk) and Model 2 (mainly sex and high-risk ACAs, only sex was significantly associated with the therapy failure risk) (Table 2). On the basis of these results, we used the original twelve variables to calculate the score (Model 3). This final score performed slightly better but was

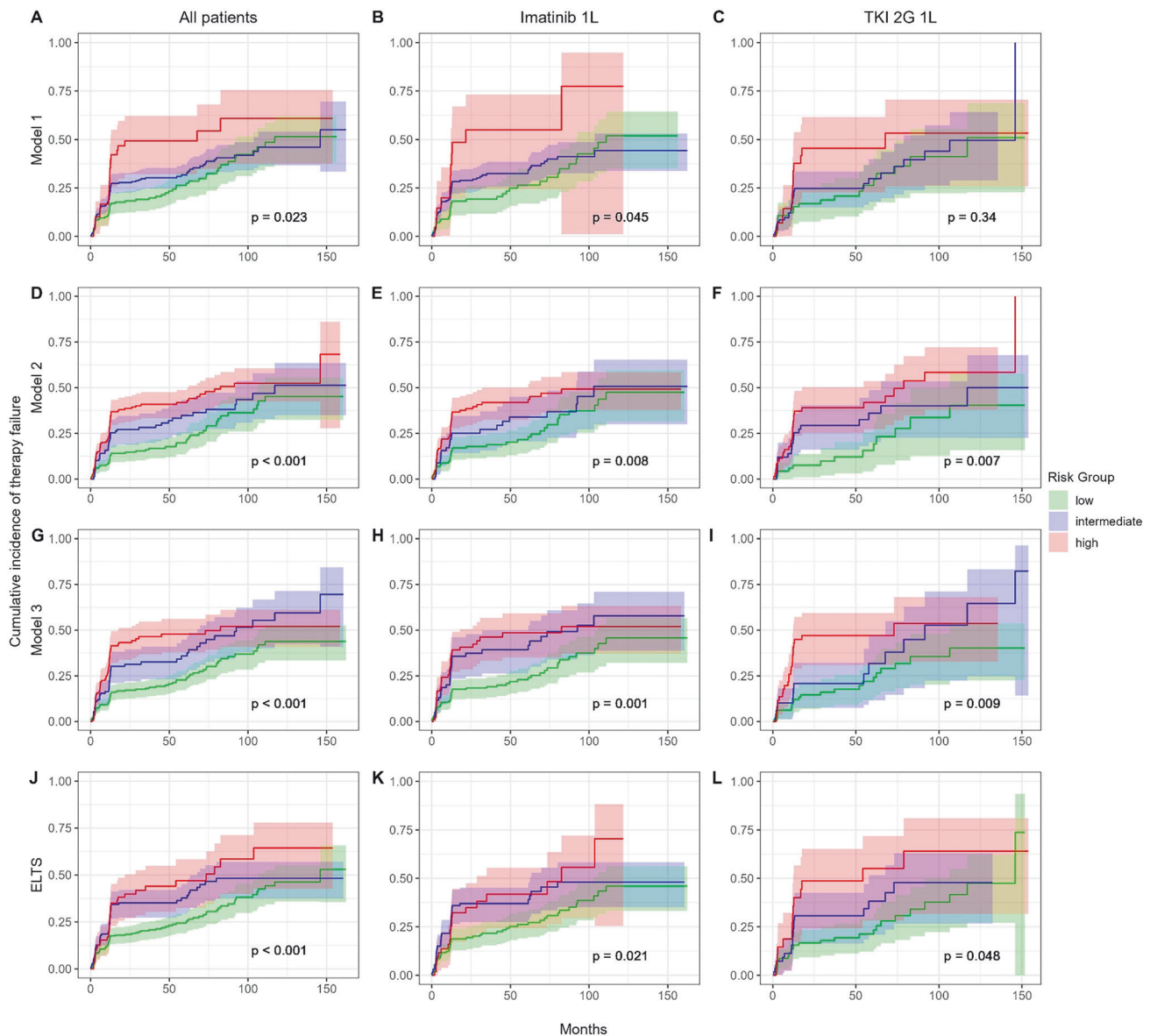


Fig. 2 Cumulative incidence of therapy failure for the three risk models and ELTS score. The cumulative incidence of treatment failure was calculated using the predictive score developed by Zhang et al. (Model 1). Model 1 was applied to the group of 601 patients selected from the CML Observatory using the selection criteria described by Zhang et al. (A) and also to the same patients after their classification into two groups: first-line (1 L) imatinib (B) and first-line 2G-TKI (C). Then, the model was reconstructed using the procedure detailed by Zhang et al. (Supplementary Information) using the six key variables identified in their study (hemoglobin, blasts, age, sex, high-risk ACAs, spleen size) (Model 2). This new model was used to calculate the risk of failure in the whole cohort (D), and in the first-line imatinib (E) and first-line 2G-TKI groups (F). In Model 3, from the initial twelve candidate variables used by Zhang et al., four variables (sex, high-risk, leukocytosis, and comorbidities) were selected to optimize the model fit. This revised model was applied to the whole cohort (G), and to the first-line imatinib (H) and first-line 2G-TKI groups (I). In parallel, the ELTS score was calculated in the whole cohort (J), first-line imatinib group (K), and first-line 2G-TKI group (L). P-values were computed with the omnibus Gray test.

not clearly superior to the traditional one, and its performance was nowhere near that of the score used in the Chinese cohort. Nevertheless, our latest approach allowed us to identify some variables of interest. Indeed, the Fine-Gray regression analysis (Model 3) identified four variables associated with the risk of therapeutic failure: sex and high-risk ACAs (like in the Chinese study), and also leukocytosis and comorbidities. This demonstrates that the weights of the variables in the statistical analysis differ depending on the patient cohort, according to their ethnic, environmental, or societal background, because the variables used here are parameters commonly considered in CML cohorts worldwide. Sex was the most discriminative factor, consistent with the higher likelihood of achieving a deep response in women

[13, 16, 17]. The reason is unclear, but it could be related to differences in the biological aging of the hematopoietic tissue [18]. The adverse prognostic impact of high-risk ACAs is well-established in blast phase-CML, and their emergence in the chronic phase constitutes a warning criterion according to the 2020 ELN recommendations [1]. Nevertheless, their influence on the TKI response prediction remains a matter of debate in CP-CML [19–21]. Here, despite the inherent limitations of real-world observational data collection, the percentage of patients harboring high-risk ACAs was higher in our cohort (5%) than in the Chinese cohort (2%) (Zhang 2024), and in an Italian cohort (2%) [22], used to evaluate the performance of the imatinib therapy failure score developed for imatinib-treated patients with CP-CML

Table 2. Multivariate analysis of covariates associated with therapy failure (Fine-Gray regression).

Model 1 (reported in Zhang et al., training cohort)			
Covariate	Regression coefficient	Hazard ratio (95% CI)	p-value
Male sex (binary)	0.1919	1.2 (1.0–1.5)	0.01
Age/100, years (continuous)	1.6160	5.0 (2.5–10.3)	<0.001
(Hemoglobin/100) ⁻² , g/L (continuous)	0.3105	1.4 (1.2–1.5)	<0.001
Blood blasts, % (continuous)	0.1087	1.1 (1.1–1.2)	<0.001
Spleen size below the costal margin, cm (continuous)	0.0671	1.1 (1.0–1.1)	<0.001
High-risk ACAs (binary)	0.5461	1.7 (1.1–2.8)	0.02
Model 2 (French cohort)			
Covariate	Regression coefficient	Hazard ratio (95% CI)	p-value
Male sex (binary)	0.5042	1.66 (1.18–2.32)	0.0032
Age/100, years (continuous)	0.3609	1.43 (0.442–4.66)	0.55
(Hemoglobin/100) ⁻² , g/L (continuous)	0.0450	1.05 (0.629–1.74)	0.86
Blood blasts, % (continuous)	0.0381	1.04 (0.968–1.11)	0.29
Spleen size below the costal margin, cm (continuous)	0.0107	1.01 (0.972–1.05)	0.59
High-risk ACAs (binary)	0.4360	1.55 (0.933–2.56)	0.091
Model 3 (French cohort)			
Covariate	Regression coefficient	Hazard ratio (95% CI)	p-value
Male sex (binary)	0.4817	1.62 (1.16–2.26)	0.0045
High-risk ACAs (binary)	0.4252	1.53 (0.947–2.47)	0.082
White blood cells, ×10E + 9/L (continuous)	0.0015	1.00 (1.00–1.00)	0.034
Comorbidities (binary)	0.2609	1.32 (0.968–1.74)	0.081

Model 1 considers the six variables selected by Zhang et al. in the Fine-Gray regression. The indicated values are the ones reported in the original study (their Table 2). Model 2 considers the same six variables, but the Fine-Gray regression coefficients were recalculated using data from our French cohort. Model 3 is the result of variable selection on a full Fine-Gray regression considering the twelve candidate variables in Zhang et al., computed using data from the French cohort.

ACAs additional chromosomal abnormalities.

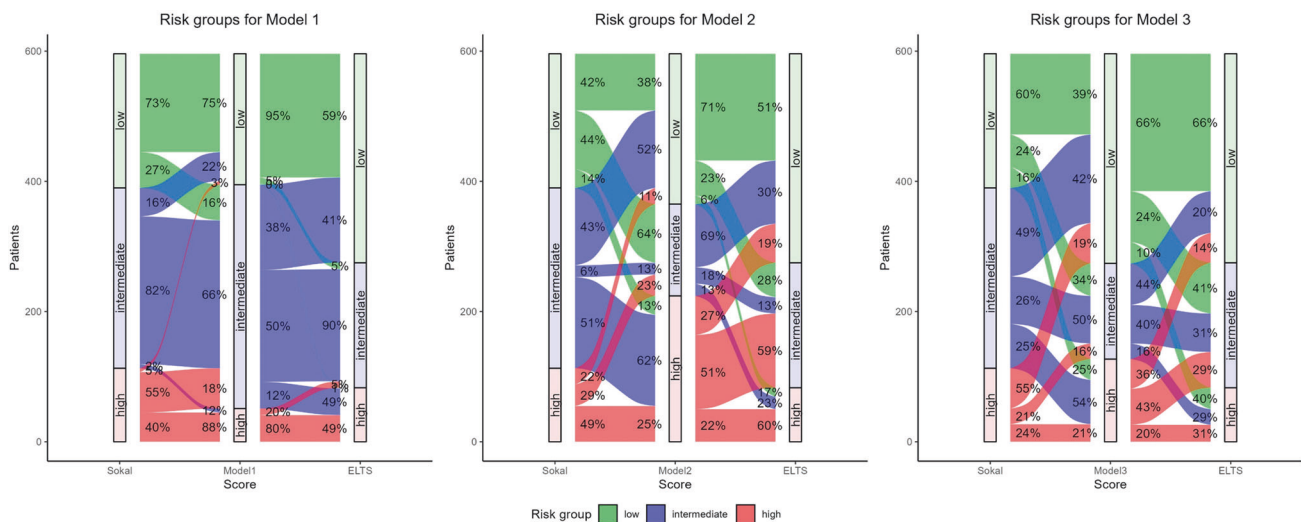


Fig. 3 Alluvial plots showing the patient distribution across the risk categories defined by the Sokal, ELTS, and three new predictive models (Model 1, 2, and 3). Each panel illustrates the concordance and discrepancies in patient risk classification between the Sokal score, ELTS score, and one of the three models (from left to right: Model 1, Model 2, Model 3). Patients are grouped into three risk categories according to each scoring system: low (green), intermediate (blue), and high (red). The flow of patients between categories is depicted by the width of the connecting bands, indicating the proportion of patients reclassified into different risk groups by the different scoring systems.

in 2022 [7]. This higher proportion may have facilitated the identification of the association between ACAs and the risk of treatment failure observed in our study, and in line with some findings [20]. The identification of peripheral leukocytosis might seem surprising, but the analysis focused specifically on TKI failure

in patients with CP-CML in whom leukocytosis has been associated with the response to imatinib [23], transcript type [24], and CD26⁺ cell [25] percentage, which are not available in our database. Comorbidities, which are not included in any CML prognostic score to date, improved the score performance.

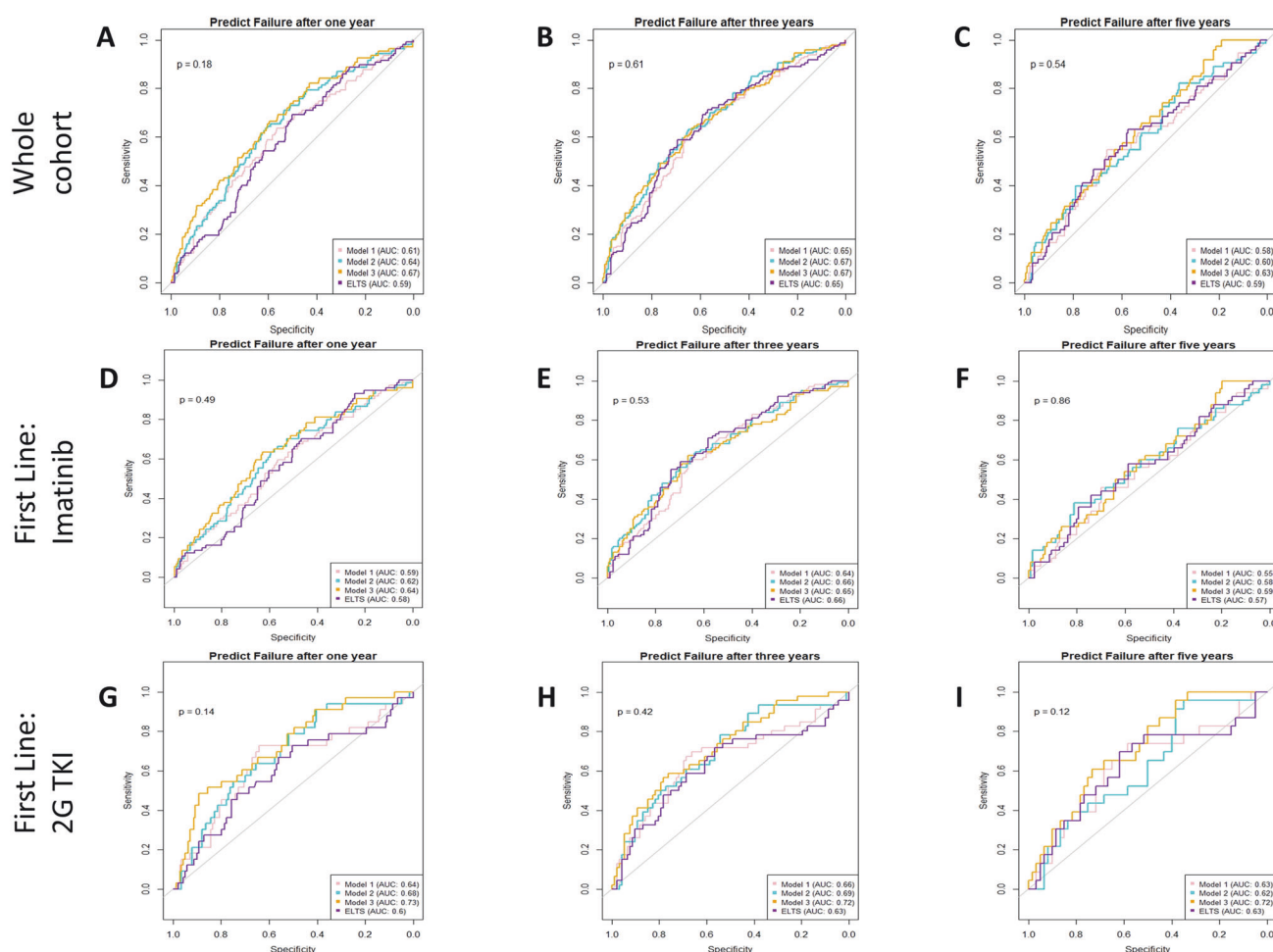


Fig. 4 Predictive performance of the three models and ELTS score. Receiver-operating characteristic (ROC) curves to compare the predictive performance of the three models and ELTS score in the whole cohort and in the first-line imatinib and first-line 2G-TKI groups. Time-dependent AUC values for therapy failure at 1 (A, D, G), 3 (B, E, H), and 5 years (C, F, I) using data from the French CML Observatory. No significant difference was detected with the omnibus Gray test (p is indicated in each graph).

Comorbidities should be taken into account when choosing a TKI [26] and are associated with an increased risk of mortality [27]; however, their impact on therapeutic efficacy/resistance through changes in patient adherence [28], drug interactions, or side effects [29] remains poorly understood. Age was a non-significant covariate, in line with the reduced influence of age [30] since TKI advent, justifying efforts to develop a new age-independent score. The earlier onset of CML may explain the identification of age in the Chinese cohort.

These results suggest that the conventional parameters available at diagnosis are insufficient to predict the risk of TKI failure in a European-type cohort and that other important factors also play a role. Some of these may be identifiable at diagnosis, such as additional genetic abnormalities [31, 32], and might influence the risk of TKI failure and consequently the predictive score [33, 34]. However, the difficulty of constructing a predictive score for the response to or failure of targeted therapies in CML might also be explained by the considerable influence of the initial treatment efficacy and tolerance on the patient trajectory and long-term therapeutic outcomes. Real-world data highlight the complexity of care pathways, and to date, there is no diagnostic criterion that allows the individual prediction of the clinical course. More specifically, intolerances, often unpredictable, can affect both dose and adherence, thereby reducing the likelihood of achieving the expected therapeutic response. In addition, complex therapeutic sequences may adversely influence the chances of

attaining an optimal outcome [35]. Thus, a major challenge in developing predictive scores of the therapeutic response lies in integrating diagnostic parameters and early follow-up data into a dynamic predictive algorithm. Moreover, this approach requires the precise and systematic collection of data throughout the patient follow-up [9]. Our findings show that, unlike in our cohort, the initial diagnostic parameters were of paramount importance in the Chinese cohort. Therefore, it would be of great interest to compare the therapeutic sequences and patient trajectories in different cohorts because these may vary significantly among countries.

Our study has several limitations. Like the study by Zhang et al., it was a retrospective analysis. The cohort size was smaller, but it is a multicentric cohort and representative of French real-world CML management. In addition, the group of patients analyzed in this study has characteristics comparable to the French cohort, and the number of patients analyzed was sufficient to assess the performance of this new score. Indeed, among the 601 patients considered, 188 (31%) experienced treatment failure. This number of events is sufficient to establish a robust survival model with ≤ 12 predictors, guaranteeing a minimum of 15 events per variable, as recommended by various studies [36, 37]. Furthermore, given the overlapping confidence intervals of the hazard ratios for the three risk groups, the differences between groups should not become substantially more pronounced, even with a larger cohort, as seen in Zhang et al. (Supplementary Fig. 1). Also, it should be noted that

Models 2 and 3 were calibrated with the same dataset used for the ROC comparisons. This introduces a risk of optimism bias due to overfitting because these scores were evaluated using data that may have implicitly influenced their construction. Conversely, the ELTS and Zhang scores were developed independently using external cohorts. Although it is unlikely that the observed score performances will become close to those of the Chinese cohort, ideally, all models should be compared using a fully independent validation cohort to avoid this bias. To clarify these aspects, meta-analysis approaches involving cohorts of different origins could be interesting, as well as strengthening international collaboration, as done by German and Italian groups before the TKI era [38].

In conclusion, our study demonstrates the limitations of directly transposing predictive models developed in ethnically or clinically distinct populations and also offers new insights that could advance personalized treatment strategies in CP-CML. By reconstructing the predictive model in a European real-world cohort, we identified distinct and clinically relevant variables (sex, leukocytosis, comorbidities, and high-risk ACAs) that may better reflect the therapeutic dynamics in this setting. These findings support the notion that resistance to TKIs results from multifactorial and context-dependent mechanisms that are only partially captured by diagnostic parameters alone. Our results are further arguments in favor of the need to develop dynamic, population-specific scoring tools that also integrate the early therapeutic responses and patient-specific trajectories. Ultimately, these tools could enable a more precise risk stratification and could contribute to improving the long-term outcomes through better-adapted TKI choices and monitoring strategies.

DATA AVAILABILITY

For original data, please contact the corresponding author, Marc G. BERGER (mberger@chu-clermontferrand.fr).

REFERENCES

- Hochhaus A, Baccarani M, Silver RT, Schiffer C, Apperley JF, Cervantes F, et al. European LeukemiaNet 2020 recommendations for treating chronic myeloid leukemia. *Leukemia*. 2020;34:966–84.
- Sokal JE, Cox EB, Baccarani M, Tura S, Gomez GA, Robertson JE, et al. Prognostic discrimination in 'good-risk' chronic granulocytic leukemia. *Blood*. 1984;63:789–99.
- Pfirschmann M, Baccarani M, Saussele S, Guilhot J, Cervantes F, Ossenkoppele G, et al. Prognosis of long-term survival considering disease-specific death in patients with chronic myeloid leukemia. *Leukemia*. 2016;30:48–56.
- Sato E, Iriyama N, Tokuhira M, Takaku T, Ishikawa M, Nakazato T, et al. The EUTOS long-term survival score predicts disease-specific mortality and molecular responses among patients with chronic myeloid leukemia in a practice-based cohort. *Cancer Med*. 2020;9:8931–9.
- Zhang XS, Gale RP, Huang XJ, Jiang Q. Is the Sokal or EUTOS long-term survival (ELTS) score a better predictor of responses and outcomes in persons with chronic myeloid leukemia receiving tyrosine-kinase inhibitors?. *Leukemia*. 2022;36:482–91.
- Saugues S, Lambert C, Daguene E, Roth-Guepin G, Huguet F, Cony-Makhoul P, et al. The initial molecular response predicts the deep molecular response but not treatment-free remission maintenance in a real-world chronic myeloid leukemia cohort. *Haematologica*. 2024;109:2893–2907.
- Zhang XS, Gale RP, Zhang MJ, Huang XJ, Jiang Q. A predictive scoring system for therapy-failure in persons with chronic myeloid leukemia receiving initial imatinib therapy. *Leukemia*. 2022;36:1336–42.
- Zhang X, Liu B, Huang J, Zhang Y, Xu N, Gale RP, et al. A predictive model for therapy failure in patients with chronic myeloid leukemia receiving tyrosine kinase inhibitor therapy. *Blood*. 2024;144:1951–61.
- Lauseker M, Hoffmann VS, Pfirschmann M. Prognostic scoring systems in chronic myeloid leukaemia. *Leukemia*. 2025;39:1046–52.
- Austin PC, Fine JP. Practical recommendations for reporting Fine-Gray model analyses for competing risk data. *Stat Med*. 2017;36:4391–400.
- Gray RJ. A Class of K-Sample Tests for Comparing the Cumulative Incidence of a Competing Risk. *Ann Stat [Internet]*. 1988 Sep 1 [cited 2025 Jun 2];16(3). Available from: <https://projecteuclid.org/journals/annals-of-statistics/volume-16/issue-3/A-Class-of-K-Sample-Tests-for-Comparing-the-Cumulative-Incidence-of-a-Competing-Risk/1176350951.full>
- DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics*. 1988;44:837.
- Kim H, Kim S, Kim HJ, Kim YK, Kwak JY, Yhim HY, et al. Comparison of frequency and sensitivity of BCR-ABL1 kinase domain mutations in Asian and white patients with imatinib-resistant chronic-phase chronic myeloid leukemia. *Clin Lymphoma Myeloma Leuk*. 2018;18:e391–9.
- Lim E, Miyamura J, Chen JJ. Racial/ethnic-specific reference intervals for common laboratory tests: a comparison among Asians, Blacks, Hispanics, and Whites. *Hawaii J Med Public Health J Asia Pac Med Public Health*. 2015;74:302–10.
- Hu S, Chen D, Xu X, Zhang L, Wang S, Jin K, et al. Targeted next-generation sequencing identifies additional mutations other than BCR::ABL in chronic myeloid leukemia patients: a Chinese Monocentric Retrospective Study. *Cancers*. 2022;14:5752.
- Branford S, Yeung DT, Ross DM, Prime JA, Field CR, Altamura HK, et al. Early molecular response and female sex strongly predict stable undetectable BCR-ABL1, the criteria for imatinib discontinuation in patients with CML. *Blood*. 2013;121:3818–24.
- Breccia M, Molica M, Colafigli G, Massaro F, Quattrocchi L, Latagliata R, et al. Prognostic factors associated with a stable MR4.5 achievement in chronic myeloid leukemia patients treated with imatinib. *Oncotarget*. 2018;9:7534–40.
- Zhang C, Hao T, Bortoluzzi A, Chen MH, Wu X, Wang J, et al. Sex-dependent differences in hematopoietic stem cell aging and leukemogenic potential. *Oncogene*. 2025;44:64–78.
- Alhuraiji A, Kantarjian H, Boddur P, Ravandi F, Borthakur G, DiNardo C, et al. Prognostic significance of additional chromosomal abnormalities at the time of diagnosis in patients with chronic myeloid leukemia treated with frontline tyrosine kinase inhibitors. *Am J Hematol*. 2018;93:84–90.
- Hehlmann R, Voskanyan A, Lauseker M, Pfirschmann M, Kalmanti L, Rinaldetti S, et al. High-risk additional chromosomal abnormalities at low blast counts herald death by CML. *Leukemia*. 2020;34:2074–86.
- Decamp M, Klein E, Godon C, Lestrangier V, Roynard P, Theisen O, et al. Cytogenetics in the management of myeloproliferative neoplasms, mastocytosis and myelodysplastic/myeloproliferative neoplasms: Guidelines from the Group Francophone de Cytogénétique Hématologique (GFCH). *Curr Res Transl Med*. 2023;71:103424.
- Ielo C, Scalzulli E, Carosino I, Pepe S, Bisegna ML, Martelli M, et al. Validation of imatinib therapy failure score (IMTF) in chronic phase chronic myeloid leukemia in real life practice. *Leuk Lymphoma*. 2023;64:2324–6.
- Lekovic D, Gotic M, Milic N, Zivoinovic B, Jovanovic J, Colovic N, et al. Predictive parameters for imatinib failure in patients with chronic myeloid leukemia. *Hematology*. 2017 Mar 22;1–7.
- Vasconcelos AP, Azevedo IF, Melo FCBC, Neves WB, Azevedo ACAC, Melo RAM. BCR-ABL1 transcript types showed distinct laboratory characteristics in patients with chronic myeloid leukemia. *Genet Mol Res*. 2017;16. <https://doi.org/10.4238/gmr16029541>.
- Culen M, Borsky M, Nemethova V, Razga F, Smejkal J, Jurcek T, et al. Quantitative assessment of the CD26+ leukemic stem cell compartment in chronic myeloid leukemia: patient-subgroups, prognostic impact, and technical aspects. *Oncotarget*. 2016;7:33016–24.
- Jabbour E, Kantarjian H. Chronic myeloid leukemia: 2025 update on diagnosis, therapy, and monitoring. *Am J Hematol*. 2024;99:2191–212.
- Saußeles S, Krauß MP, Hehlmann R, Lauseker M, Proetel U, Kalmanti L, et al. Impact of comorbidities on overall survival in patients with chronic myeloid leukemia: results of the randomized CML Study IV. *Blood*. 2015;126:42–9.
- Breccia M, Latagliata R, Stagno F, Luciano L, Gozzini A, Castagnetti F, et al. Charlson comorbidity index and adult comorbidity evaluation-27 scores might predict treatment compliance and development of pleural effusions in elderly patients with chronic myeloid leukemia treated with second-line dasatinib. *Haematologica*. 2011;96:1457–61.
- Rosti G, Brummendorf TH, Gjertsen BT, Giraldo-Castellano P, Castagnetti F, Gambacorti-Passerini C, et al. Impact of age and comorbidities on the efficacy and tolerability of bosutinib in previously treated patients with chronic myeloid leukemia: results from the phase 4 BYOND study. *Leukemia*. 2024;38:126–35.
- Latagliata R, Breccia M, Carosino I, Cesini L, De Benedittis D, Mohamed S, et al. Clinical results according to age in patients with chronic myeloid leukemia receiving imatinib frontline: The younger, the later, the worse?. *Eur J Haematol*. 2018;101:578–84.
- Schönfeld L, Rinke J, Hinze A, Nagel SN, Schäfer V, Schenk T, et al. ASXL1 mutations predict inferior molecular response to nilotinib treatment in chronic myeloid leukemia. *Leukemia*. 2022;36:2242–9.
- Nteliopoulos G, Bazeos A, Claudiani S, Gerrard G, Curry E, Szydlo R, et al. Somatic variants in epigenetic modifiers can predict failure of response to imatinib but not to second-generation tyrosine kinase inhibitors. *Haematologica*. 2019;104:2400–9.

33. Perusini MA, Žáčková D, Kim T, Pagnano K, Pavlovsky C, Ježíšková I, et al. Mutations in myeloid transcription factors and activated signaling genes predict chronic myeloid leukemia outcomes. *Blood Adv.* 2024;8:2361–72.
34. Guerineau H, Cayuela J, Dulucq S, Tran Quang V, Tarfi S, Gricourt G, et al. Mutation of Epigenetic Regulators At Diagnosis Is An Independent Predictor Of Tyrosine Kinase Inhibitor Treatment Failure In Chronic Myeloid Leukemia: A Report From the RESIDiAG study. *Am J Hematol.* 2025;100:507–10.
35. Bosi GR, Fogliatto LM, Costa TEV, Grokoski KC, Pereira MP, Bugs N, et al. What happens to intolerant, relapsed or refractory chronic myeloid leukemia patients without access to clinical trials?. *Hematol Transfus Cell Ther.* 2019;41:222–8.
36. Harrell, FE. Regression modeling strategies: with applications to linear models, logistic and ordinal regression, and survival analysis [Internet]. Cham: Springer International Publishing; 2015 [cited 2025 Apr 14]. (Springer Series in Statistics). Available from: <https://link.springer.com/10.1007/978-3-319-19425-7>
37. Hair JF, Black WC, Babin BJ, Anderson RE. Multivariate data analysis. 7. Auflage, Pearson new internat. ed. Harlow: Pearson; 2014. 1 p. (ProQuest Ebook Central).
38. Hasford J, Baccarani M, Hehlmann R, Anseri H, Tura S, Zuffa E. Interferon-alpha and hydroxyurea in early chronic myeloid leukemia: a comparative analysis of the Italian and German chronic myeloid leukemia trials with interferon-alpha [letter]. *Blood.* 1996;87:5384–91.

ACKNOWLEDGEMENTS

The CML Observatory Steering Committee would like to thank the national academic expert group Fi-LMC (chairman: Dr Franck Nicolini) for their support in recognizing the CML Observatory as the real-life database for CML in France. We would also like to thank the investigators from the centers not included in this study (Drs. Gabriel Etienne (Bordeaux), Shanti Ame (Strasbourg), Karine Soulier-Guérin (Vichy), Anne-Blandine Boutin (Contamine-sur-Arve), Aude Charbonnier (Marseille)) for joining the CML Observatory and contributing to its development.

AUTHOR CONTRIBUTIONS

MGB and CB designed the study, participated in the data analysis, and participated in the writing of the manuscript. SM carried out all statistical analyses and participated in the writing of the manuscript. BL organized the extraction of data from the CML Observatory, participated in the data analysis, and participated in the writing. BL and SM contributed equally to the study. TL extracted the data and participated in the analysis of the results. HJA, CD, GRG, FH, LR, AP, MM, AP, LL, VP, MEB, PR, GD, HM, and MGB are investigators in the partner centers of the CML Observatory, who included patients, recorded data. JB and AK participated in the data analysis. DH and MGB developed the computer database; DH maintains it and is responsible for data security and data protection. All authors validated the submitted manuscript.

COMPETING INTERESTS

The authors declare no competing interests.

ETHICAL DECLARATIONS

Consent to participate by signed informed consent was obtained from all patients included in the study. The CML Observatory has been authorized (authorization no. 914456) by CNIL, the French data protection authority, in accordance with its ethical standards and with the 1964 Helsinki declaration and its later amendments or comparable ethical standards. It is promoted by the Clermont-Ferrand University Hospital, that is the coordinating center.

ADDITIONAL INFORMATION

Supplementary information The online version contains supplementary material available at <https://doi.org/10.1038/s41375-025-02703-6>.

Correspondence and requests for materials should be addressed to Marc G. Berger.

Reprints and permission information is available at <http://www.nature.com/reprints>

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Open Access This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by-nc-nd/4.0/>.

© The Author(s) 2025