Allergy

WILEY | EAACI

ORIGINAL ARTICLE OPEN ACCESS

Drug Allergy, Insect Sting Allergy, and Anaphylaxis

Development of Betalactam-Predictor: A Clinical Decision Tool for Delabeling Low-Risk Betalactam Allergy Patients. Initial Validation in Penicillin Allergy

Marina Labella[1,2] | Rafael Nuñez[2] | Inmaculada Doña[1,2] | Julia Rodríguez de Guzmán[1,2,3] | Esther Moreno[4] | Lene Heise Garvey[5,6] | Jose Julio Laguna[7] | Annick Barbaud[8] | Patrizia Bonnadona[9] | Jonas Bredtoft Boel[10] | Holger Mosbech[5] | Giovanna Sfriso[11] | Mariana Castells[12,13] | Elizabeth Phillips[14] | María José Torres[1,2,3]

[1]Allergy Unit, Hospital Regional Universitario de Málaga, Málaga, Spain | [2]Allergy Research Group, Instituto de Investigación Biomédica de Málaga y Plataforma en Nanomedicina- IBIMA Plataforma BIONAND, Málaga, Spain | [3]Departamento de Medicina, Facultad de Medicina, Universidad de Málaga, Málaga, Spain | [4]Allergy Service, University Hospital of Salamanca, Salamanca, Spain Institute for Biomedical Research of Salamanca (IBSAL), Salamanca, Spain | [5]Allergy Clinic, Department of Dermatology and Allergy, Herlev and Gentofte Hospital, University of Copenhagen, Copenhagen, Denmark | [6]Department of Clinical Medicine, Copenhagen University, Copenhagen, Denmark | [7]Allergy Unit, Allergo- Anaesthesia Unit, Faculty of Medicine, Hospital Central de la Cruz Roja, Alfonso X El Sabio University, Madrid, Spain | [8]Sorbonne Université, INSERM, Institut Pierre Louis D'epidémiologie et de Santé Publique, AP- HP. Sorbonne Université, Hôpital Tenon, Service de Dermatologie et Allergologie, Paris, France | [9]Allergy Unit, Ospedale San Bortolo, Vicenza, Italy | [10]Department of Clinical Microbiology, Herlev and Gentofte Hospital, University of Copenhagen, Copenhagen, Denmark | [11]Allergy Unit, Azienda Ospedaliera Universitaria Integrata Verona, Verona, Italy | [12]Division of Allergy and Clinical Immunology, Department of Medicine, Brigham and Women's Hospital, Boston, Massachusetts, USA | [13]Department of Medicine, Harvard Medical School, Boston, Massachusetts, USA | [14]Department of Infectious Diseases, Vanderbilt University Medical Centre, Nashville, Tennessee, USA

Correspondence: Elizabeth Phillips (elizabeth.j.phillips@vanderbilt.edu) | María José Torres (mjtorresj@gmail.com)

Received: 20 August 2025 | **Revised:** 11 December 2025 | **Accepted:** 15 December 2025

Handling Editor: Robyn E. O'Hehir

Keywords: anaphylaxis | challenge tests | drug allergy

ABSTRACT

Background: A label of betalactam (BL) allergy is estimated in around 10% of the population in their medical records. Second-line choices carry significant negative consequences, including reduced efficacy, effectiveness, and safety. This study aimed to develop a new highly specific score constructed by selecting variables assisted by artificial intelligence to identify low-risk BL-allergic patients.

Methods: In this study, derivation and validation of the BL-predictor score were performed on a retrospective cohort of 2207 patients who underwent penicillin allergy testing at Málaga University Hospital (Spain). The development of the BL-predictor encompassed expert drafting and a two-step variable selection process consisting of univariate analysis and variable filtering, followed by stepwise logistic regression with resampling. To assess the efficiency, a multicentric retrospective external validation was performed in 4261 patients from six populations: Salamanca and Madrid, Spain; Nashville, United States of America; Verona, Italy; Paris, France; and Copenhagen, Denmark.

Results: The definitive questionnaire consisted of eight items and risk points were computed from the logistic regression model as follows: +1 for reactions after first dose or in less than 1 h (ITEM-1), +2 for anaphylaxis (ITEM-2); +1 for previous reaction with the culprit (ITEM-3); −1 for resolution in > 24 h (ITEM-4); +2 for spontaneous resolution (ITEM-5); −2 for unknown symptoms

Marina Labella, Rafael Núñez and Inmaculada Doña contributed equally to this study and are considered joint first authors. Elizabeth Phillips and María José Torres should be considered joint last authors.

(ITEM-6); −2 for reaction occurred > 5 years (ITEM-7), and −1 for another reported drug allergy (ITEM-8). After establishing a threshold of ≤0 points to classify individuals with low risk, internal validation showed a specificity of 86% and a negative predictive value (NPV) of 83%. Overall multicenter external validation showed a specificity of 93%, which implies a 25% increase in specificity compared to the previously published BL decision tool.

Conclusion: This score would simplify diagnostic procedures in low-risk patients, enabling rapid delabeling, potentially in non-specialty settings, and reducing diagnostic costs and the negative consequences associated with incorrect antibiotic allergy labels.

1 | Introduction

Betalactam (BL) antibiotics are the drugs most commonly involved in allergic reactions. It is estimated that approximately 10% of the population has a documented history of BL allergy in their medical records, totaling millions of people globally [1]. A label of BL allergy carries significant negative consequences for the treatment of the underlying infection, including reduced efficacy, effectiveness, and safety because of the use of more costly second-line antibiotics. Ultimately, this leads to increased antibiotic resistance, longer hospital stays, and increased mortality [2–9]. This is especially significant as less than 5%–30% of adults with an allergy label are truly allergic [3, 10]. Given the widespread use of BL antibiotics, it is essential to identify methods for removing incorrect labels (delabeling) of allergy to BL antibiotics. Current safe strategies indicate that over 90% of children and adults can be delabeled with a single-dose oral challenge [11, 12].

However, a major obstacle to successful delabeling is the difficulty of diagnosing allergy to BL antibiotics. The diagnostic work-up starts with a clinical history, which, particularly in the case of a remote reaction, may not be helpful to verify key details. Diagnosis of BL allergy has included in vitro tests that are not highly sensitive, are not available at all centers for all BL drugs, or are not fully validated. Skin testing, which lacks 100% of negative predictive value, and if done in the context of low-risk reactions with a very low pre-test probability, could lead to false positives and a low positive predictive value [13–17]. Therefore, the drug provocation test (DPT), otherwise known as oral challenge (OC), is the gold standard, which has been done in a specialized setting, and it is time-consuming and costly. In essence, the diagnostic work-up is complex, long (it can take several days), not risk-free, and inefficient [10, 18].

Over recent decades, there has been a growing interest in creating methods to risk-stratify patients on the basis of clinical history, permitting the adaptation of diagnostic strategies to the risk and optimizing investigations in terms of efficiency and resource usage, while maintaining patient safety. In this regard, risk stratification tools may help to identify low-risk patients where the pre-test probability of true penicillin allergy is extremely low and where direct DPT would be preferred.

Several recent guidelines have addressed the need to define risk categories [10, 19–20]. Although there is a general consensus on identifying high-risk individuals [10, 19–20], the definition of low-risk remains controversial. Although the recent EAACI position paper on DPT agrees that low-risk reactions are characterized by mild cutaneous non-immediate reactions (NIRs), such as maculopapular exanthema (MPE) with no danger signs and resolving in less than 1 week [12, 21–23], there is still a lack of consensus concerning other types of reactions. Additionally, it suggests that delayed urticaria and MPE lasting more than 7 days are considered intermediate risk, and unknown reactions fall into the low-intermediate risk category [19], deemed unsuitable for direct DPT. Nevertheless, a recent study showed that MPE, regardless of duration and delayed urticaria, as well as other skin lesions such as delayed angioedema and diverse categories like unknown childhood reactions and syncope in remote reactions without any other associated symptoms, were safe for direct single-dose DPT and should be considered as low-risk [24].

Artificial intelligence (AI) breakthroughs have driven novel healthcare improvements and have been used to extract data from electronic records to classify subjects according to severity and confirmed diagnosis [25–27]. Focusing on automated diagnosis/risk assessment for drug hypersensitivity, several efforts have been made [28, 29], including an artificial neural network to predict BL-hypersensitivity on the basis of clinical history [30] and an AI-assisted BL allergy delabeling in perioperative adverse reactions evaluation [31].

On the other hand, simple scores and questionnaires that can be easily implemented in penicillin allergy evaluation have become very popular, such as the 1–1-1 criterion for identifying high-risk [25] and the PEN-FAST clinical decision rule for identifying low-risk patients [25, 32]. PEN-FAST [25] prioritizes key historical factors including the time since the last reaction, the presence of severe symptoms such as anaphylaxis or angioedema, and the need for medical intervention. It has been validated in American (USA) and Australian populations with a negative predictive value of 96.3%. However, despite its usefulness, its applicability may be limited in children and other populations where the prevalence of penicillin allergy and phenotypes may vary [33]. In fact, when PEN-FAST was tested in European populations, it struggled to accurately predict relapse of immediate skin hypersensitivity or delayed maculopapular exanthema, with 28.6% and 38.4% of patients misclassified, respectively [34]. These difficulties underscore the need to develop and adapt new risk stratification scoring systems that can be reliably adapted and applied across different clinical contexts and populations globally.

The aim of this study was to develop a new highly specific score constructed by selecting variables assisted by AI to identify low-risk BL-allergic patients. This score would then expedite and simplify diagnostic procedures in low-risk patients, achieving rapid and scalable delabeling, potentially in non-specialty settings, which will lead to reduced diagnostic costs and the negative consequences associated with incorrect antibiotic allergy labels.

2 | Materials and Methods

2.1 | Design

The initial derivation and validation of the BL-predictor were done on a retrospective cohort of patients and controls 18 years and older evaluated with a history of hypersensitivity reactions to penicillins between 1985 and 2024 in Malaga University Hospital (Spain). Allergological work-up was performed a minimum of 4 weeks after the reaction, including ST, and if negative, DPT. Allergic cases were those with positive results on ST or, in those with negative ST, positive results on DPT with the culprit; controls were those with confirmed tolerance to the culprit BL in DPT.

STs were done by skin prick test (SPT) and, if negative, by an intradermal test (IDT), as previously published [10, 35–36], using daily-prepared solutions. Reagents were PPL (Penicilloyl poly-L-lysine) and MDM (minor determinant mixture) (Allergopen; Allergopharma-JGKG, Reinbek, Germany) until 2006; and BP-OL (0.04 mg/mL) and MD (minor determinant) (0.5 mg/mL) (Benzylpenicilloyl-octa-l-lysine and benzylpenilloate, DAP-Diater, Madrid, Spain) afterward, amoxicillin (DAP, Diater laboratories, Madrid, Spain), clavulanic acid (DAP, Diater laboratories, Madrid, Spain), cloxacillin (Normon, Madrid, Spain), and piperacillin-tazobactam (Stada, Barcelona, Spain). United States' population used AllerQuest LLC major determinant (ALK-Abelló), and the clinic prepared minor determinant with ampicillin 25 mg/mL, benzyl penicillin 1000 and 10,000 IU/mL, and any other relevant culprit penicillin (e.g., piperacillin-tazobactam). Reading was done after 20 min and considered positive: (i) in the SPT, if a wheal larger than 3 mm surrounded by erythema appeared, with a negative response to the control saline; and (ii) in the intradermal test, if the diameter of the wheal area that was marked initially increased more than 3 mm and was surrounded by erythema [35]. Positive data are expressed as the mean diameter recorded by measuring the largest and smallest diameters at right angles to each other [10, 35].

DPT was done in a blind, placebo-controlled manner with the culprit drug at incremental doses (25%–25%–50% of a full therapeutic single dose, if history suggested anaphylaxis the first dose was divided into 10% and 15%), with a minimum 30-min interval between each, reaching a therapeutic dose. Since 2023, patients reporting mild non-immediate reactions with no alarm signs were challenged with a direct full single dose.

Patients were monitored during DPT procedures and for 2 h after the last dose of DPT. Complete cardiopulmonary resuscitation equipment was immediately available. If tolerated, NIR DPT was followed by a 48 h washout period and then a 2-day treatment course at home (1 dose every 12 h) [36]. In those cases who reported unknown penicillins, DPT was performed with amoxicillin.

The external validation was done in retrospective patients from different hospitals: Salamanca University Hospital (Salamanca—Spain), Hospital de la Cruz Roja (Madrid—Spain), Vanderbilt University Medical Center (Nashville, TN—United States of America), University Hospital of Verona (Verona—Italy), Hôpital Tenon (Paris—France), and Gentofte Hospital (Copenhagen—Denmark), as shown in Table 1.

The variables included in the database were age, sex, history of atopy, familial history of atopy, and patient-reported episodes suggestive of hypersensitivity to a BL antibiotic, including symptoms, specific BL implicated, number of suggestive reactions, treatments, latency (time between administration of the BL and onset of reaction), and date of the episode. Reported reactions with missing data, due to the inability to recall them at the allergological assessment, were codified as unknown. Consequently, unknown entries were considered a valid value, that is, an informative answer in the anamnesis.

The study was conducted in accordance with the principles of the Declaration of Helsinki and approved by the institutional review panel. Informed consent was obtained from all patients for all the diagnostic procedures. Approval from the ethics committee or institutional review board was obtained at each institution.

2.2 | BL-PREDICTOR Development and Validation

A first version of the clinical decision tool BL-predictor was designed, inspired by published criteria [19, 26], and a panel of experts in BL-hypersensitivity, composed of members of the EAACI project "Delabeling penicillin, a patient-driven tool", including Prof. Torres, Prof. Phillips, and Prof. Castells as chairs, and Ph.D. Labella as secretary. In a first phase, seven questions were evaluated with a univariate analysis to address their discrimination potential. Next, this preliminary approach was followed by an optimization phase (variable selection) in two steps. A graphical abstract of this workflow is depicted in Figure 1.

To identify new potential predictors, in the first step of the optimization phase, reported reactions variables from our database were filtered according to: 1—discriminating potential between BL-hypersensitivity patients (medium and high risk) and low-risk individuals (univariate analysis significance threshold of $p < 0.05$), 2—prevalence (frequency greater than 10% in either risk level populations), and 3—non-existent or weak collinearity with any of the 7 expert's committee items (Crammer's $V < 0.20$) (Table S6).

In the second step, to obtain the optimal combination of individual predictors selected by step 1, including experts' proposed items, an AI-powered multivariate analysis was conducted. A machine learning (ML) based strategy was developed to generate an optimized logistic regression model using stepwise variable selection and bootstrapping resampling. MASS R package's stepwise logistic regression, with up to 1000 forward or backwards steps, automatically trains a model with the best combination of candidate variables according to the Akaike Information Criterion (AIC), AIC measures the quality of a given model relative to any other. On top of this, bootstrapping resampling (10,000 iterations) implemented in caret's R package ensures that the model selected, by AIC, in the stepwise regression is robust and closer to generalization; each iteration consists of a random sampling of 75% of the data and an execution of the stepwise logistic regression algorithm.

TABLE 1 | Description of population characteristics.

		Malaga		Spain[a]		Spain[b]		Italy[c]		France[d]		USA[e]		Denmark[f]		
BL-Hypersensitivity?		Yes 37% (811)	No 63% (1396)	Yes 52% (118)	No 48% (109)	Yes 29% (592)	No 61% (2033)	Yes 49% (65)	No 51% (69)	Yes 15% (22)	No 85% (126)	Yes 2% (20)	No 98% (996)		Yes 14% (99)	No 86% (604)
Diagnosis method	ST	37%	—	61%	—	88%	—	77%	—	82%	—	75%	—		0%	—
	DPT	63%		39%		12%		23%		18%		25%			100%	
Culprit Allergy label	Amoxicillin	33%	30%	46%	36%	46%	33%	92%	94%	50%	57%	40%	13%	Aminopenicillins	33%	21%
	AX-CLV	57%	33%	46%	32%	34%	20%	3%	4%	45%	41%	5%	7%			
	Penicillin	7%	34%	11%	36%	16%	47%	—	—	—	—	35%	80%	Isoxazolylpenicillins	8%	6%
	Penicillin G	0%	0%	—	—	—	—	—	—	—	—	—	—			
	Penicillin V	0%	1%	—	—	—	—	—	—	—	—	—	—	Penicillin G/V	13%	17%
	Cloxacillin	0%	1%	—	—	—	—	—	—	—	—	—	—			
	Ampicillin	1%	0%	—	—	—	—	—	—	5%	2%	5%	—	Penicillins with BLI	6%	6%
	Piperacillin-Tazobactam	1%	1%	—	—	—	—	—	—	—	—	15%	—	Unknown penicillins	39%	50%
Latency	Immediate	59%	16%	79%	63%	73%	42%	80%	54%	68%	28%	25%	7%		4%	4%
	Non-Immediate	27%	51%	21%	7%	4%	10%	20%	46%	32%	63%	50%	82%		78%	68%
	Unknown	14%	33%	—	30%	23%	48%	—	—	—	9%	25%	11%		18%	28%

Note: Immediate reactions were considered those that occurred within 1 h.
Abbreviations: DPT, Drug provocation test; NA, not available; ST, Skin tests.
[a]Hospital Central de la Cruz Roja (Madrid – Spain).
[b]Salamanca University Hospital (Salamanca – Spain).
[c]University Hospital of Verona (Verona – Italy).
[d]Hôpital Tenon (Paris – France).
[e]Vanderbilt University Medical Center (Nashville, TN – United States of America).
[f]Gentofte Hospital (Copenhagen – Denmark).

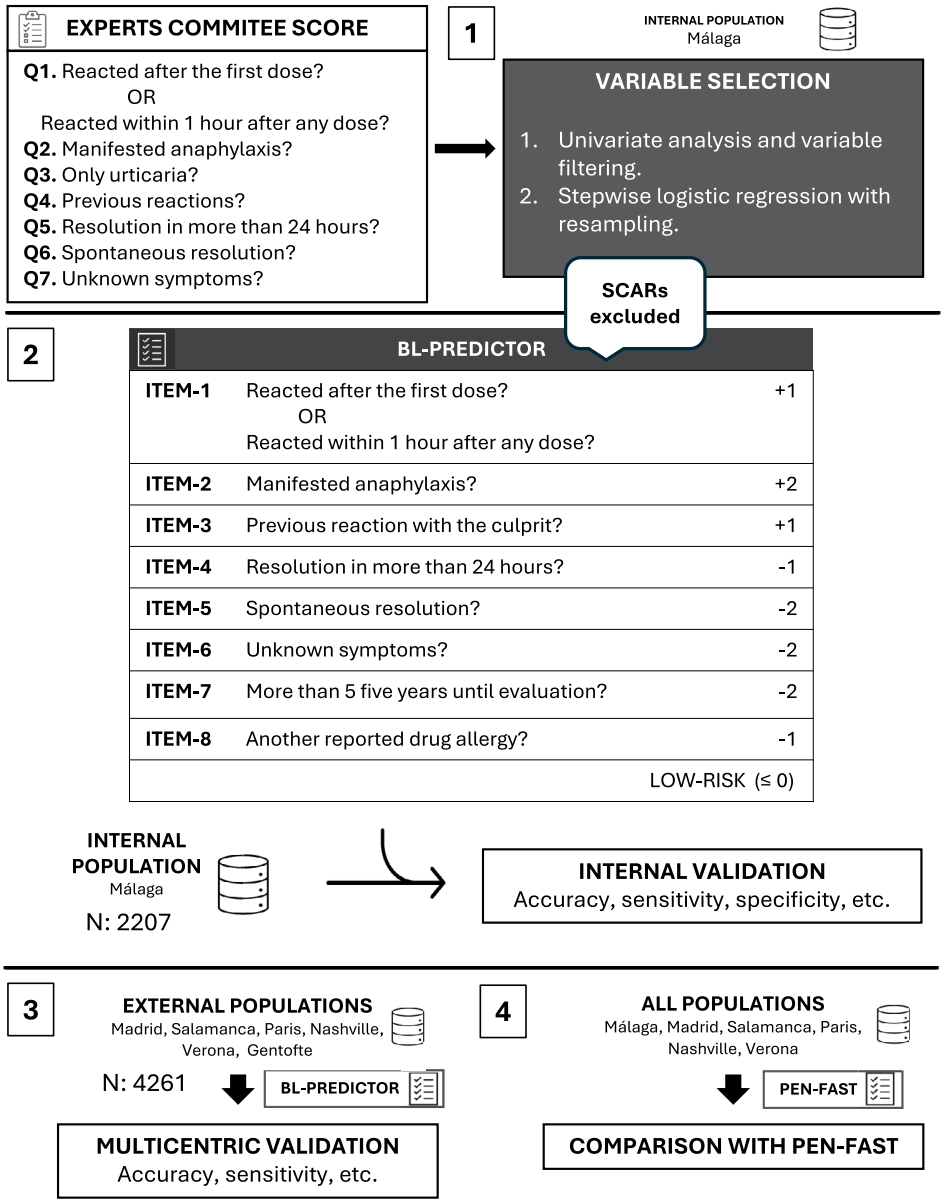

FIGURE 1 | (1) Development of BL-predictor, encompassing the experts draft and the two-steps variable selection. (2) Definitive BL-predictor questionnaire and internal validation. (3) External multicenter validation of BL-predictor and comparison against another low-risk stratification score (PEN-FAST). SCARs (Severe Cutaneous Adverse Reactions). SCARs are defined by presenting desquamation, fever, pustules, multiple lymphadenopathies, blisters, mucosal lesions, and/or eosinophilia.

Therefore, this ML strategy produces a model built with an optimal set of variables to include in the definitive BL-predictor, removing residual collinearity and estimating final variables' contribution to risk classification. Training was conducted with R packages caret and MASS [37, 38]. Lastly, to simplify BL-predictor calculations, risk (negative or positive) points were assigned to each item in the BL-predictor by rounding its coefficients estimated in the optimized logistic regression model trained in optimization's phase step 2 (Table S2). Univariate and multivariate analyses were performed as described in the statistical methods section. Patients presenting with symptoms compatible with Severe Cutaneous Adverse Reactions (SCARs) such as desquamation, fever, pustules, multiple lymphadenopathies, blisters, mucosal lesions, and/or eosinophilia did not fill in the questionnaire since SCARs are considered a contraindication to DPT [38].

Following development, the BL-predictor was validated with the internal population, assessing its performance and suitability. First, the risk score was calculated for all individuals by aggregating their BL-predictor points. Next, after establishing a threshold of ≤ 0 points to classify individuals with low risk of suffering BL-hypersensitivity, classification performance was estimated to measure BL-predictor power to identify low-risk subjects. In the same manner, to assess the efficiency of BL-predictor as a clinical decision tool, a multicentric external validation was performed with six populations from five different countries (Table S1). External databases were imported with *readxl*, *haven* R packages and processed with the *base* R (version 4.5.0) package [39–45], collecting the data necessary to compute the BL-predictor risk score and to conduct a descriptive analysis. Finally, to compare BL-predictor with PEN-FAST, the PEN-FAST score was computed for populations included in this

TABLE 2 | Univariate analysis of the clinical history variables selected in the first step of the optimization phase obtained by Malaga's population.

Univariate analysis						
		No/low-risk (1396%–63%)		Yes/medium or high-risk (811%–37%)		
Confirmed betalactam allergy?		***N***	**Freq (%)**	***N***	**Freq (%)**	***p* (χ2 test)[a]**
Sex	Female	926	66	481	59	< 0.001
	Male	469	34	330	41	
History of an additional drug allergy (outside of BL allergy)		121	9	35	4	< 0.001
Skin symptoms	Angioedema	276	20	260	32	< 0.001
Time from initial reaction to evaluation	> 5 years	319	23	80	10	< 0.001

[a]All variables were categorical or categorized, χ2 test was applied to compare its separation potential between low-risk and medium/high-risk classes.

study, except for the Danish population. Performance was measured with the *base*, *stats*, *pROC, PropCIs*, and *caret* R packages [37, 46–47].

2.3 | Statistical Methods

For the univariate analysis, variables were tested according to their nature and distribution. First, for numerical variables, homoscedasticity was tested by the Breusch–Pagan test, and then Student's *t*-test or Welch's *t*-test was carried out. To measure the effect size, Cohen's *D* was used; effect size for numerical variables was interpreted following Cohen's guidelines: 0–0.20 (none or weak), 0.21–0.50 (small), 0.51–0.80 (medium), and 0.81–1 (large) [48].

Second, categorical, or categorized, variables were tested with the Chi-squared test. Multicollinearity between categorical variables was studied with Crammer's *V*; all variables passing significance and prevalence filters were categorical or categorized. Crammer's *V* interpretation was done as follows: *V* in [0,0.1) (no associated), *V* in [0.1,0.2) (weak association), *V* in (0.2,0.3) (moderate association), *V* in [0.4,0.6) (relatively strong association), *V* in [0.6,0.8) (strong association), *V* [0.8,1] (very strong association) [49].

Descriptive exploratory analysis was conducted with *base*, *rstatix*, *vtable, statmod*, and *Hmisc* R packages [50–53]. For categorical variables, relative frequencies were computed. For numerical variables, the mean and the standard deviation, or the median and the interquartile range, were calculated. Odds ratios of BL-predictor items in Tables 2 and S2 were computed in R with the ad/bc formula.

3 | Results

3.1 | Expert Committee Draft Validation

To find variables suitable to complement the initial questionnaire, a filtering procedure was conducted. Clinical history variables were selected from the database if they had statistically significant differences ($p < 0.05$), prevalence over 10% in any of the two risk levels, and low collinearity with the aforementioned items. Candidate variables (Table 2) included sex, another reported drug allergy, manifestation of angioedema in any reported episode, and an interval between the reaction and the clinical evaluation greater than 5 years, with the latter showing the larger classification potential ($X^2 = 335.128$). Next, an ML approach was applied to construct a stepwise logistic regression model with bootstrapping resampling. This algorithm selects the most informative combination of variables, thus optimizing the discrimination between low-risk and high-risk subjects and estimating the contribution to the prediction of each variable selected. The final logistic regression model is detailed in Table S2. Variable selection discarded Q3 (urticaria as sole symptom) and added two new items: interval reaction-evaluation greater than 5 years (new ITEM-7) and history of another reported drug allergy (ITEM-8). All eight variables present in this model were statistically significant; three of them were associated with an increase in risk (ITEMS 1–3), whereas the rest were related to a decreased probability of suffering BL-hypersensitivity (ITEMS 4–8). Therefore, the definitive questionnaire consisted of eight items (Figure 1.2). Risk points were computed from the logistic regression model (Figure 1.2) and were assigned as follows: +1 points to reactions after the first dose or in less than 1 h after any dose (ITEM-1) and to more than one reported episode (ITEM-3), +2 points were awarded to the presence of anaphylaxis (ITEM-2). Anaphylaxis was defined with the following criteria: presence of skin (warmth, erythema, urticaria, pruritus, and/or angioedema) and respiratory symptoms (wheezing, dyspnea, oxygen desaturation, and/or cough) or presence of respiratory symptoms and/or skin symptoms with cardiovascular symptoms (tachycardia, presyncope, syncope, and/or hypotension).

According to logistic regression coefficients (Table S2): minus 2 points were assigned to spontaneous resolution (ITEM-5), Interval Reaction-Evaluation (> 5y) (ITEM-7), and unknown symptoms (ITEM-6) items. Besides, delayed (> 24 h) resolution (ITEM-4) and history of another reported drug allergy (ITEM-8) –1 point score was assigned (Figure 1). BL-predictor's score frequencies in the different populations are shown in Figure 2.

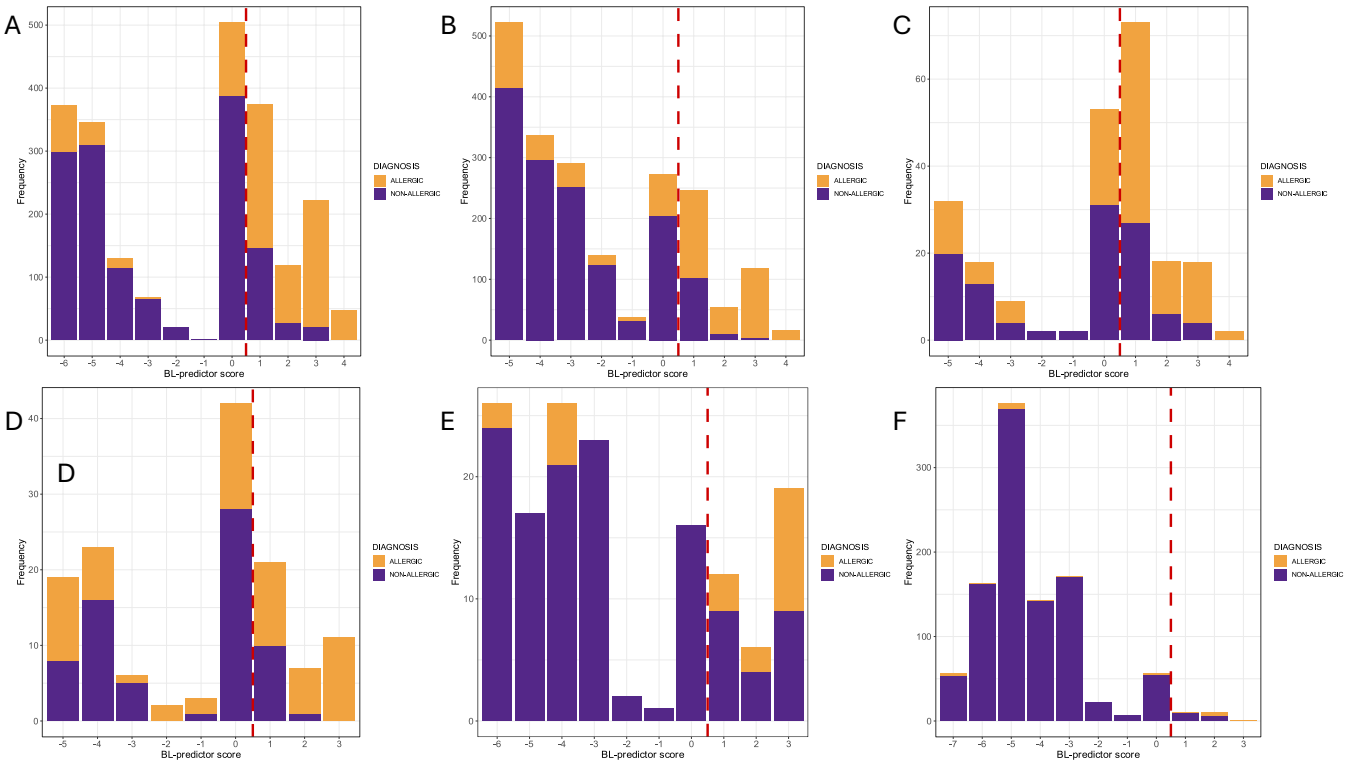

FIGURE 2 | BL-predictor's score frequencies in the different populations. (A) Spain-Málaga. (B) Spain-Salamanca. (C) Spain-Madrid. (D) Italy-Verone. (E) Paris-France. (F) Nashville-USA. Allergic and non-allergic fractions are colored orange and blue, respectively. BL-predictor's threshold is represented as a dashed read line.

3.2 | Malaga Population Demographic and Clinical Data

The internal cohort included 2207 patients, 1407 (63.75%) female, mean age 44.5 years old (SD 16). A total of 811 (37%) were confirmed as allergic: 297 had positive skin results, and 514 reacted to DPT. The most common allergy label was AXCLV ($n = 925$), and the most frequent phenotype was urticaria/angioedema ($n = 1234$). The remaining demographic and clinical data from Malaga's population used to develop the tool are summarized in Table 3, and descriptive statistics and test results are shown for high-medium-risk and low-risk subjects. Significant differences ($p < 0.001$) appeared in the number of episodes, culprit, type of reaction, and latency variables. Patients categorized as high-risk had more reported episodes (1.3 ± 0.53 versus 1.1 ± 0.33), involving histories compatible with anaphylaxis (42% vs. 6%), with a more immediate latency (59% vs. 16%), and less frequently caused by penicillin (7% vs. 34%). A female preponderance was also seen in high-risk patients (Female: 66% vs. Male: 59%).

3.3 | Experts Committee Items Univariate Analysis

Experts' committee items' ability to discriminate low-risk subjects was analyzed. Candidate items' univariate analysis is shown in Table 4. All seven items exhibited notable statistically significant differences ($p < 0.001$) and thus great potential for differentiation.

3.4 | BL-Predictor Internal Validation

To assess the performance of BL-predictor, the score was computed for each subject in our database ($n = 2207$) to classify them as low or medium/high-risk patients; classification was evaluated with confirmed diagnosis information. In Table 5 different low-risk classification performance metrics are shown. BL-predictor stood out as a specificity-centered tool, obtaining a specificity of 86%, correctly classifying at least three-quarters of the population (accuracy = 80% and AUC = 0.78) with the questionnaire.

3.5 | BL-Predictor Multicentric Validation

Demographic and clinical data summaries of the external populations can be found in Table S1, each BL-predictor item univariate analysis in these populations is shown in Table S3. BL-predictor score was computed for each patient in the complete joined multicentric population of suspected BL-hypersensitivity patients' database ($N = 4261$), individuals were classified as low or medium-high risk patients. In Table 5, different low-risk classification performance metrics are shown for the external population. Overall, the specificity analysis indicated good discrimination for the BL-predictor score in all external databases, including 66.06% (56.36%–74.85%) for Madrid (Spain), 91.81% (90.27%–93.18%) for Salamanca (Spain), 84.06% (73.26%–91.76%) for Italy, 82.54% (74.77%–88.72%) for France, 98.49% (97.53%–99.15%) for the USA, and 92.88% (90.53%–94.8%) for Denmark (Table 6).

TABLE 3 | Malaga's population demographics univariate analysis.

		No/low-risk (1396%–63%)		Yes/high-risk (811%–37%)			
Betalactam allergy?		**Mean**	**SD**	**Mean**	**SD**	***p*[a] (Cohen's *D*)**	
Age at diagnosis (years)		45	17	44	15	0.219 (0.052)	
Number of episodes		1.1	0.33	1.3	0.53	< 0.001 (0.302)	
		***N* (Yes)**	**Freq (%)**	***N* (Yes)**	**Freq**	***p*[b]**	
Sex	Female	926	66	481	59	< 0.001	
	Male	469	34	330	41		
Culprit	Amoxicillin	412	30	267	33	0.104	< 0.001
	AX-CLV	461	33	464	57	< 0.001	
	Penicillin	477	34	54	7	< 0.001	
	Penicillin G	4	0	5	0	0.658	
	Penicillin V	8	1	10	0	0.341	
	Cloxacillin	13	1	4	0	0.319	
	Ampicillin	6[C]	0	8	1	0.162	
	Piperacillin-Tazobactam	15	1	11	1	0.546	
Type of Reaction	Urticaria/Angioedema	836	60	398	49	< 0.001	< 0.001
	Anaphylaxis	80	6	337	42	< 0.001	
	Exanthema	68	5	23	3	0.020	
	Respiratory	23	2	6	1	0.111	
	Non-suggestive/Other	240	17	38	5	< 0.001	
	Unknown	145	10	7	1	< 0.001	
Latency	Immediate	217	16	480	59	< 0.001	
	Non-Immediate	710	51	221	27		
	Unknown	463	33	110	14		

[a]Welch's *t*-test.
[b]$\chi 2$ test or Fisher's exact test.
[C]Low-risk individuals reported anaphylaxis frequency is overestimated by self-reporting bias.

3.6 | Comparison Performance of Initial and Overall Population

Multicentric performance almost matched the previous performance obtained for Malaga's population, showing higher accuracy (83.26% vs. 80.02%) and specificity (92.65% vs. 85.89%), but lower sensitivity (49.45% vs. 69.91%). In terms of accuracy, Salamanca, Paris, Copenhagen, and Nashville (USA) had results greater than the internal population, whereas Madrid and Verona populations underachieved. The highest specificity value arose with Nashville's population, and 5 of 7 populations studied surpassed 80%. Besides, none of the external populations excelled in sensitivity, maximum of 68.18, with three populations falling below 50%. In addition, Verona's population obtained an accuracy of 64% but did detect 83% of low-risk individuals, although significant differences were detected only for items 1 and 2, and items 4 and 6 could not be answered with the information available. In addition, BL-predictor predictions for Nashville and Copenhagen populations had similar behavior; in both cases, a high specificity, 98% and 93% respectively, and low sensitivity, 30% and 13%, was achieved. Both populations had a low confirmed BL allergy to BL-labeled allergic subject ratio. Both the Nashville and Copenhagen populations did not show significant differences in at least one of the items that increase BL-predictor score and thus raise sensitivity; these were item 3 for Nashville and item 1 for Copenhagen. The BL-predictor score was decreased indiscriminately for true BL-allergic patients in both populations.

3.7 | BL-Predictor vs. PEN-FAST

In internal validation, PEN-FAST had a specificity of 79.15% (76.93%–81.26%) and an NPV of 76.52% (74.25%–78.69%). In overall multicenter validation, specificity was 67.79% (66%–69.53%) and NPV 88.43% (86.99%–89.77%). PEN-FAST obtained

TABLE 4 | BL-predictor experts panel candidate items univariate analysis in Malaga's population.

	No/low-risk ($N=1396$)			Yes/medium or high-risk ($N=811$)			
Betalactam allergy?	**True**	**False**	**Unknown#**	**True**	**False**	**Unknown#**	***p*[a]**
Q1 – Reaction after the first dose? OR Immediate reaction? (< 1H)	269 (19%)	658 (47%)	463 (33%)	543 (67%)	158 (19%)	110 (14%)	< 0.001
Q2 – Anaphylaxis?	80 (6%)	1171 (84%)	145 (10%)	337 (58%)	419 (42%)	7 (1%)	< 0.001
Q3 – Urticaria alone?	188 (23%)	616 (76%)	7 (1%)	491 (35%)	760 (54%)	145 (10%)	< 0.001
Q4 – Recurrence?	116 (8%)	1280 (92%)	0 (0%)	176 (78%)	635 (22%)	0 (0%)	< 0.001
Q5 – Resolution in > 24H?	344 (25%)	142 (10%)	910 (65%)	77 (9%)	191 (24%)	543 (67%)	< 0.001
Q6 – No treatment needed?	258 (19%)	1077 (77%)	61 (4%)	55 (7%)	726 (90%)	30 (4%)	< 0.001
Q7 – Unknown reaction?	145 (10%)	1251 (90%)	0 (0%)	7 (1%)	804 (99%)	0 (0%)	< 0.001

[a]$\chi 2$ test or Fisher's exact test #.

TABLE 5 | Performance metrics for BL-predictor in Málaga's population (internal validation) and multicenter joined populations (external validation).

BL-predictor performance metrics	**BL-predictor internal validation ($N=2207$)**	**Multicenter joined performance ($N=4261$ [916 positive cases])**
Accuracy (%)	80.02 (78.29–81.67)	83.26 (82.21–84.47)
Sensitivity (%)	69.91 (66.63–73.05)	49.45 (46.17–52.74)
Specificity (%)	85.89 (83.95–87.67)	92.65 (91.71–93.51)
Positive predictive value (%)	74.21 (70.96–77.28)	64.81 (61.14–68.35)
Negative predictive value (%)	83.09 (81.06–84.99)	87.00 (85.85–88.09)
AUC	0.79 (0.77–0.79)	0.76 (0.74–0.76)
Cohen's Kappa	0.56 (0.56–0.57)	0.46 (0.46–0.46)

an improvement in sensitivity (+12%) in the external populations at the cost of specificity (−28%) (Table S4). PEN-FAST individual external population results are shown in Table S5.

4 | Discussion

BL-predictor is a clinical decision tool refined by AI approaches that has demonstrated high specificity (internal validation 86% and external validation 93%) for delabeling low-risk penicillin labels. In recent years, validated clinical decision tools [25, 54–55] have gained increasing interest for optimization practices, although published tools cannot be assumed to replicate across different populations. This finding is important because BL-predictor has been shown to be useful in heterogeneous populations, indicating that it could be used as a universal delabeling tool. Overall, multicenter external validation showed an increase of 25% in specificity compared to previously published BL decision tools. According to the BL-predictor's performance in the Malaga population used to develop the tool, the questionnaire allowed correct classification in 86% of non-allergic patients, with only 14% misclassified. Regarding this, 14% of non-allergic patients were misclassified; a quarter of them were computed as medium-high risk because ITEM-2 (anaphylaxis) was marked. Those patients likely reported symptoms involving at least two organs, but with normal vital signs, since sometimes patients experience subjective symptoms that can be mistaken for genuine allergic symptoms. Therefore, BL-predictor offers rapid and effective low-risk stratification, enabling alternative diagnostic approaches that could make the allergological assessment significantly more efficient while maintaining safety. For example, in Málaga's population, BL-predictor would allow 75.38% of the subjects (scoring 0 or less) to avoid ST, whereas the PEN-FAST could enable 60.99% of the individuals (scoring 2 or less) to evade ST. Concerning safety, 70% of the hypersensitivity to BL antibiotics patients were deemed as medium-high risk, implying a 30% false negative rate; however, only 0.4% of the BL-hypersensitivity patients reporting systemic reactions

TABLE 6 | Stratified multicenter BL-predictor performance.

BL-predictor performance metrics (stratified)						
	Spain[a]	Spain[b]	Italy[c]	France[d]	USA[e]	Denmark[f]
Accuracy (%)	64.32 (57.71–70.55)	80.67 (78.88–82.36)	64.18 (55.44–72.27)	80.41 (73.09–86.47)	97.15 (95.93–98.08)	81.65 (78.59–84.44)
Sensitivity (%)	62.71 (53.33–71.44)	53.55 (49.44–57.62)	43.08 (30.85–55.96)	68.18 (45.13–86.14)	30 (11.89–54.28)	13.13 (7.18–21.41)
Specificity (%)	66.06 (56.36–74.85)	91.81 (90.27–93.18)	84.06 (73.26–91.76)	82.54 (74.77–88.72)	98.49 (97.53–99.15)	92.88 (90.53–94.8)
Positive predictive value (%)	66.67 (57.09–75.33)	72.87 (68.43–77)	71.79 (55.13–85)	40.54 (24.75–57.9)	28.57 (11.28–52.18)	23.21 (12.98–36.42)
Negative predictive value (%)	62.07 (52.59–70.91)	82.79 (80.85–84.61)	61.05 (50.5–70.89)	93.69 (87.44–97.43)	98.59 (97.65–99.23)	86.70 (83.85–89.23)
AUC	0.64 (0.58–0.64)	0.78 (0.76–0.78)	0.66 (0.58–0.66)	0.67 (0.59–0.67)	0.64 (0.54–0.64)	0.53 (0.49–0.55)
Cohen's Kappa	0.29 (0.28–0.29)	0.49 (0.49–0.49)	0.27 (0.27–0.28)	0.40 (0.38–0.41)	0.28 (0.27–0.28)	0.07 (0.07–0.08)

[a]Hospital Central de la Cruz Roja (Madrid – Spain).
[b]Salamanca University Hospital (Salamanca – Spain).
[c]University Hospital of Verona (Verona – Italy).
[d]Hôpital Tenon (Paris – France).
[e]Vanderbilt University Medical Center (Nashville, TN – United States of America).
[f]Gentofte Hospital (Copenhagen – Denmark).

were classified as low-risk by BL-predictor. Hence, BL-predictor detects most BL-allergy labeled patients with severe systemic reactions as medium-high risk (97%), therefore reliably identifying those at risk of developing severe reactions at DPT where this procedure would be appropriately avoided.

Furthermore, BL-predictor was tested in a multicenter international validation study and achieved a global accuracy of 91%, correctly stratifying 95% and 58% of low and medium-high risk patients, respectively, in line with what was observed in the development of the tool. However, individual prediction of the external populations showed disparities in the accuracy of the BL-predictor, possibly because of the differences in the phenotypes and diagnostic approaches in each population (Table S1). In this sense, maybe some populations do not include anaphylaxis patients since they are not confirmed as allergic because skin testing was negative and DPT was not performed.

As noted above, there is a substantial heterogeneity among the different populations of individuals with reported BL-hypersensitivity reactions, as several factors linked to risk in the Malaga population were not associated with confirmed BL allergy diagnosis in others. This has been previously documented in BL allergy delabeling tools, including PEN-FAST and validation studies [25, 34, 46]. To compare PEN-FAST and BL-predictor potential, the PEN-FAST score was computed. BL-predictor stood up for being more accurate in all the populations studied, mean accuracy 78% versus 62%, particularly for achieving a greater specificity in all the predictions, mean specificity difference of +20%. This capacity for the BL-predictor to risk-stratify and identify true low-risk individuals suggests that this could have a positive impact and allow scale-up of high-throughput delabeling outside of the allergy specialty setting.

In some populations, MPE prevalence is higher, maybe because of two factors: general practitioners diagnosing IR by specific IgE and not referring, and recruitment bias. In these populations, the long duration of evolution and resolution of MPE could be thought of as a factor that increases the risk of becoming sensitized, although ITEM 4 (Resolution in more than 24 h −1) is considered 'protective' for being allergic. For this reason, an additional item (duration more than 7 days) has been added to the PEN-FAST score (+ score) [34] with the aim of improving the performance for those with NIRs. In our study, we have included two centers in which NIRs are more common (38% and 68% were NIRs, respectively), maintaining good results for NPVs (94% and 87%, respectively).

In conclusion, the BL-predictor was developed as a refined tool designed to achieve specificity in identifying patients appropriate for DPT for penicillin allergy delabeling. Our study shows that it is an accurate point-of-care tool to perform risk stratification on patients with unverified BL allergy. By reliably identifying low-risk patients who can safely go straight to DPT, BL-predictor has the potential to allow scalable BL delabeling outside of the specialty setting. This will reduce the overall global burden of BL-hypersensitivity diagnosis. Although these results are promising, validation across centers internationally to confirm its generalizability in different treatment settings is still needed. The development and validation of the BL-Predictor underscore the advantages and emerging opportunities associated with using sophisticated AI-driven tools that harness recent technological progress and increasing public trust in AI [56].

Author Contributions

M.L., I.D., R.N., M.C., E.P., and M.J.T. designed the study and coordinated the work of the rest of the authors. M.L., I.D., and M.J.T. recruited the study individuals from the Málaga population for the internal validation and obtained clinical data. R.F. performed the statistical analysis and AI methodology. E.M., J.J.L., L.H.G., J.R.G., E.P., P.B., G.S., A.B., J.B.B., H.M., and M.J.T. recruited study individuals and obtained clinical data for their respective populations for the external validation. M.L., I.D., R.N., E.P., and M.J.T. wrote the manuscript, tables, and figures, which were reviewed by the rest of the authors.

Acknowledgements

We thank A. Soria and J. Castagna for their help in recruiting the Paris-Tenon population.

Funding

This Project was supported by the European Academy of Allergy and Clinical Immunology (EAACI) under the EAACI Committee, 43309.

Conflicts of Interest

The authors declare no conflicts of interest.

Data Availability Statement

The authors have nothing to report.

References

1. I. Dona, M. J. Torres, G. Celik, E. Phillips, L. K. Tanno, and M. Castells, "Changing Patterns in the Epidemiology of Drug Allergy," *Allergy* 79, no. 3 (2024): 613–628.

2. M. Castells, D. A. Khan, and E. J. Phillips, "Penicillin Allergy," *New England Journal of Medicine* 381, no. 24 (2019): 2338–2351.

3. K. G. Blumenthal, J. G. Peter, J. A. Trubiano, and E. J. Phillips, "Antibiotic Allergy," *Lancet* 393, no. 10167 (2019): 183–198.

4. K. G. Blumenthal, K. Kuper, L. T. Schulz, et al., "Association Between Penicillin Allergy Documentation and Antibiotic Use," *JAMA Internal Medicine* 180, no. 8 (2020): 1120–1122.

5. D. R. MacFadden, A. LaDelfa, J. Leen, et al., "Impact of Reported Beta-Lactam Allergy on Inpatient Outcomes: A Multicenter Prospective Cohort Study," *Clinical Infectious Diseases: An Official Publication of the Infectious Diseases Society of America* 63, no. 7 (2016): 904–910.

6. J. R. Pano-Pardo, E. Moreno Rodilla, S. Cobo Sacristan, et al., "Management of Patients With Suspected or Confirmed Antibiotic Allergy: Executive Summary of Guidelines From the Spanish Society of Infectious Diseases and Clinical Microbiology (SEIMC), the Spanish Society of Allergy and Clinical Immunology (SEAIC), the Spanish Society of Hospital Pharmacy (SEFH) and the Spanish Society of Intensive Medicine and Coronary Care Units (SEMICYUC)," *Journal of Investigational Allergology & Clinical Immunology* 33, no. 2 (2023): 95–101.

7. M. J. Torres, N. F. Adkinson, Jr., J. C. Caubet, et al., "Controversies in Drug Allergy: Beta-Lactam Hypersensitivity Testing," *Journal of Allergy and Clinical Immunology in Practice* 7, no. 1 (2019): 40–45.

8. K. G. Blumenthal, N. Lu, Y. Zhang, Y. Li, R. P. Walensky, and H. K. Choi, "Risk of Meticillin Resistant Staphylococcus Aureus and *Clostridium difficile*in Patients With a Documented Penicillin Allergy:

Population Based Matched Cohort Study," *BMJ* 361, no. 27 (2018): k2400.

9. E. Macy and R. Contreras, "Health Care Use and Serious Infection Prevalence Associated With Penicillin "Allergy" in Hospitalized Patients: A Cohort Study," *Journal of Allergy and Clinical Immunology* 133, no. 3 (2014): 790–796.

10. A. Romano, M. Atanaskovic-Markovic, A. Barbaud, et al., "Towards a More Precise Diagnosis of Hypersensitivity to Beta-Lactams—An EAACI Position Paper," *Allergy* 75, no. 6 (2020): 1300–1315.

11. M. Labella, J. de Rodriguez Guzman, P. Diez-Echave, et al., "Direct Single-Dose Drug-Provocation Test Is Safe for Delabelling Penicillin Low-Risk Reactions in Adults," *Allergy* 80, no. 16 (2025): 3127–3139.

12. A. Prieto, C. Munoz, G. Bogas, et al., "Single-Dose Prolonged Drug Provocation Test, Without Previous Skin Testing, Is Safe for Diagnosing Children With Mild Non-Immediate Reactions to Beta-Lactams," *Allergy* 76, no. 8 (2021): 2544–2554.

13. C. Mayorga, G. Celik, P. Rouzaire, et al., "In Vitro Tests for Drug Hypersensitivity Reactions: An ENDA/EAACI Drug Allergy Interest Group Position Paper," *Allergy* 71, no. 8 (2016): 1103–1134.

14. C. Mayorga, G. E. Celik, M. Pascal, et al., "Flow-Based Basophil Activation Test in Immediate Drug Hypersensitivity. An EAACI Task Force Position Paper," *Allergy* 79, no. 3 (2024): 580–600.

15. A. Ariza, C. Mayorga, G. Bogas, et al., "Detection of Serum-Specific IgE by Fluoro-Enzyme Immunoassay for Diagnosing Type I Hypersensitivity Reactions to Penicillins," *International Journal of Molecular Sciences* 23, no. 13 (2022): 6992.

16. J. A. Cespedes, R. Fernandez-Santamaria, G. Bogas, et al., "Lipopolysaccharides in Combination With Amoxicillin Increases Basophil Activation Test Sensitivity to Amoxicillin IgE-Mediated Hypersensitivity," *Allergy* 79, no. 9 (2024): 2537–2542.

17. J. A. Cespedes, R. Fernandez-Santamaria, A. Ariza, et al., "Diagnosis of Immediate Reactions to Amoxicillin: Comparison of Basophil Activation Markers CD63 and CD203c in a Prospective Study," *Allergy* 78, no. 10 (2023): 2745–2755.

18. I. Dona, L. Guidolin, G. Bogas, et al., "Resensitization in Suspected Penicillin Allergy," *Allergy* 78, no. 1 (2023): 214–224.

19. A. Barbaud, L. H. Garvey, M. Torres, et al., "EAACI/ENDA Position Paper on Drug Provocation Testing," *Allergy* 79, no. 3 (2024): 565–579.

20. K. G. Blumenthal, E. S. Shenoy, C. A. Varughese, S. Hurwitz, D. C. Hooper, and A. Banerji, "Impact of a Clinical Guideline for Prescribing Antibiotics to Inpatients Reporting Penicillin or Cephalosporin Allergy," *Annals of Allergy, Asthma & Immunology: Official Publication of the American College of Allergy, Asthma, & Immunology* 115, no. 4 (2015): 294–300.

21. R. Confino-Cohen, Y. Rosman, K. Meir-Shafrir, et al., "Oral Challenge Without Skin Testing Safely Excludes Clinically Significant Delayed-Onset Penicillin Hypersensitivity," *Journal of Allergy and Clinical Immunology: In Practice* 5, no. 3 (2017): 669–675.

22. J. C. Caubet, L. Kaiser, B. Lemaitre, B. Fellay, A. Gervaix, and P. A. Eigenmann, "The Role of Penicillin in Benign Skin Rashes in Childhood: A Prospective Study Based on Drug Rechallenge," *Journal of Allergy and Clinical Immunology* 127, no. 1 (2011): 218–222.

23. S. Fransson, J. B. Boel, H. F. Mosbech, L. K. Poulsen, S. Ruff, and L. H. Garvey, "Safe De-Labeling of Patients at Low Risk of Penicillin Allergy in Denmark," *International Archives of Allergy and Immunology* 183, no. 6 (2022): 640–650.

24. M. Labella, "Direct Single-Dose Drug Provocation Test Is Safe for Delabelling Penicillin Low-Risk Reactions in Adults," (In Press).

25. J. A. Trubiano, S. Vogrin, K. Y. L. Chua, et al., "Development and Validation of a Penicillin Allergy Clinical Decision Rule," *JAMA Internal Medicine* 180, no. 5 (2020): 745–752.

26. H. Boyd and A. F. Santos, "Novel Diagnostics in Food Allergy," *Journal of Allergy and Clinical Immunology* 155, no. 2 (2025): 275–285.

27. D. Lisik, R. Basna, T. Dinh, et al., "Artificial Intelligence in Pediatric Allergy Research," *European Journal of Pediatrics* 184, no. 1 (2024): 98.

28. R. Nunez, I. Dona, and J. A. Cornejo-Garcia, "Predictive Models and Applicability of Artificial Intelligence-Based Approaches in Drug Allergy," *Current Opinion in Allergy and Clinical Immunology* 24, no. 4 (2024): 189–194.

29. J. M. Inglis, S. Bacchi, A. Troelnikov, W. Smith, and S. Shakib, "Automation of Penicillin Adverse Drug Reaction Categorisation and Risk Stratification With Machine Learning Natural Language Processing," *International Journal of Medical Informatics* 156 (2021): 104611.

30. E. M. Moreno, V. Moreno, E. Laffond, et al., "Usefulness of an Artificial Neural Network in the Prediction of Beta-Lactam Allergy," *Journal of Allergy and Clinical Immunology: In Practice* 8, no. 9 (2020): 2974–2982.

31. M. Jiang, A. Lam, L. Lam, et al., "Artificial Intelligence and the Potential for Perioperative Delabeling of Penicillin Allergies for Neurosurgery Inpatients," *British Journal of Neurosurgery* 39, no. 1 (2025): 40–43.

32. V. Sabato, F. Gaeta, R. L. Valluzzi, A. Van Gasse, D. G. Ebo, and A. Romano, "Urticaria: The 1-1-1 Criterion for Optimized Risk Stratification in Beta-Lactam Allergy Delabeling," *Journal of Allergy and Clinical Immunology: In Practice* 9, no. 10 (2021): 3697–3704.

33. A. M. Copaescu, S. Vogrin, G. Shand, M. Ben-Shoshan, and J. A. Trubiano, "Validation of the PEN-FAST Score in a Pediatric Population," *JAMA Network Open* 5, no. 9 (2022): e2233703.

34. J. Castagna, F. Chasset, J. E. Autegarden, et al., "Assessing Delayed Penicillin Hypersensitivity Using the PENFAST+ Score," *Frontiers in Allergy* 4 (2023): 1302567.

35. K. Brockow, A. Romano, M. Blanca, J. Ring, W. Pichler, and P. Demoly, "General Considerations for Skin Test Procedures in the Diagnosis of Drug Hypersensitivity," *Allergy* 57, no. 1 (2002): 45–51.

36. A. Barbaud, M. Weinborn, L. H. Garvey, et al., "Intradermal Tests With Drugs: An Approach to Standardization," *Frontiers in Medicine* 7 (2020): 156.

37. M. Kuhn, "Building Predictive Models in R Using the Caret Package," *Journal of Statistical Software* 28, no. 5 (2008): 1–26.

38. W. N. Venables, *Modern Applied Statistics With S*, Fourth ed. (Springer, 2002).

39. H. Wickham and J. Bryan, "readxl: Read Excel Files," (2023), https://readxl.tidyverse.org.

40. "Tidyverse," https://github.com/tidyverse/readxl.

41. H. Wickham, E. Miller, and D. Smith, "haven: Import and Export 'SPSS', 'Stata' and 'SAS' Files," (2023) R Package Version 2.5.4.

42. "Tidyverse," https://github.com/tidyverse/haven.

43. "WizardMac," https://github.com/WizardMac/ReadStat.

44. "Tidyverse," https://haven.tidyverse.org.

45. R Core Team, "R: A Language and Environment for Statistical Computing," (2021) R Foundation for Statistical Computing: Vienna, Austria, https://www.R-project.org/.

46. X. Robin, N. Turck, A. Hainard, et al., "pROC: An Open-Source Package for R and S+ to Analyze and Compare ROC Curves," *BMC Bioinformatics* 12, no. 17 (2011): 77.

47. "PropCIs: Various Confidence Interval Methods for Proportions," (2018) R Package Version 0.3–0.

48. J. Cohen, *Statistical Power Analysis for the Behavioral Sciences*, 2nd ed. (Erlbaum, 1988).

49. L. Rea and R. A. Parker, *Designing and Conducting Survey Research* (Jossey-Boss, 1992).

50. N. Huntington-Klein, "vtable: Variable Table for Variable Documentation," (2024) R Package Version 1.4.8, https://nickch-k.github.io/vtable/.

51. A. Kassambara, "rstatix: Pipe-Friendly Framework for Basic Statistical Tests," (2023) R Package Version 0.7.2, https://rpkgs.datanovia.com/rstatix/.

52. J. F. Harrell, "Hmisc: Harrell Miscellaneous," (2025) R Package Version 5.2–3, https://github.com/harrelfe/hmisc.

53. G. K. Smyth, "statmod: Statistical Modeling," (2003) R Package Version 1.5.0.

54. F. Cox, S. Vogrin, R. P. Sullivan, et al., "Development and Validation of a Cephalosporin Allergy Clinical Decision Rule," *Journal of Infection* 90, no. 6 (2025): 106495.

55. F. Stehlin, S. Vogrin, E. Mitri, G. A. C. Isabwe, J. A. Trubiano, and A. M. Copaescu, "International Validation of the SULF-FAST Risk-Stratification Tool for Sulfonamide Antibiotic Allergy," *JAMA Network Open* 8, no. 7 (2025): e2519113.

56. J. C. Rojas, M. Teran, and C. A. Umscheid, "Clinician Trust in Artificial Intelligence: What Is Known and How Trust Can be Facilitated," *Critical Care Clinics* 39, no. 4 (2023): 769–782.

Supporting Information

Additional supporting information can be found online in the Supporting Information section. **Table S1:** Multicenter demographic summary. **Table S2:** Optimization phase stepwise Logistic Regression model. **Table S3:** Multicentric BL-predictor items univariate analysis. **Table S4:** PEN-FAST performance for MALAGA's population and the joined multicenter population. **Table S5:** Stratified multicenter PEN-FAST performance. **Table S6:** Collinearity analysis of BL-predictor's candidate variables.