


Understanding Severe Asthma Through Small and Big Data in Spanish Hospitals: The PAGE Study

Melero Moreno C^{1,2}, Almonacid Sánchez C³ , Bañas Conejero D⁴, Quirce S^{5,6}, Álvarez Gutiérrez FJ⁷, Cardona V⁸, Sánchez-Herrero MG⁴, Soriano JB^{6,9,10,11}, on behalf of the PAGE Study Group*

¹Hospital Universitario La Princesa, Madrid, Spain

²Hospital Universitario 12 de Octubre, Madrid, Spain

³Hospital Universitario de Toledo, Toledo, Spain

⁴Specialty Care Medical Department, GlaxoSmithKline, Madrid, Spain

⁵Hospital Universitario La Paz, IdiPAZ, Madrid, Spain

⁶CIBER of Respiratory Diseases (CIBERES), Madrid, Spain

⁷Hospital Universitario Virgen del Rocío, Sevilla, Spain

⁸Hospital Universitario Vall d'Hebron, Barcelona, Spain

⁹Hospital Universitario de La Princesa, Madrid, Spain

¹⁰Facultad de Medicina, Universidad Autónoma de Madrid, Madrid, Spain

¹¹Instituto de Salud Carlos III (ISCIII), Madrid, Spain

*See Online Appendix for a full list of collaborators.

J Investig Allergol Clin Immunol 2023; Vol. 33(5): 373-382

doi: 10.18176/jiaci.0848

■ Abstract

Background: Data on the prevalence of severe asthma (SA) are limited. Electronic health records (EHRs) offer a unique research opportunity to test machine learning (ML) tools in epidemiological studies. Our aim was to estimate the prevalence of SA among asthma patients seen in hospital asthma units, using both ML-based and traditional research methodologies. Our secondary objective was to describe patients with nonsevere asthma (NSA) and SA over a follow-up of 12 months.

Methods: PAGE is a multicenter, controlled, observational study conducted in 36 Spanish hospitals and split into 2 phases: a cross-sectional phase for estimation of the prevalence of SA and a prospective phase (3 visits in 12 months) for the follow-up and characterization of SA and NSA patients. A substudy with ML was performed in 6 hospitals. Our ML tool uses EHRead technology, which extracts clinical concepts from EHRs and standardizes them to SNOMED CT.

Results: The prevalence of SA among asthma patients in Spanish hospitals was 20.1%, compared with 9.7% using the ML tool. The proportion of SA phenotypes and the features of patients followed up were consistent with previous studies. The clinical predictions of patients' clinical course were unreliable, and ML found only 2 predictive models with discriminatory power to predict outcomes.

Conclusion: This study is the first to estimate the prevalence of SA in hospitalized asthma patients and to predict patient outcomes using both standard and ML-based research techniques. Our findings offer relevant insights for further epidemiological and clinical research in SA.

Key words: Severe asthma. Prevalence. Big data. Machine learning. Natural language processing. Predictive models.

■ Resumen

Antecedentes: Los datos sobre la prevalencia del asma grave (SA) son limitados. La implantación de las historias clínicas electrónicas (EHR) ofrece una oportunidad única de investigación con tecnologías de aprendizaje máquina (ML) en los estudios epidemiológicos. El objetivo fue estimar la prevalencia del SA entre los pacientes atendidos en las unidades de asma hospitalarias, utilizando el ML como la metodología de investigación tradicional. Los objetivos secundarios fueron describir los pacientes con asma no grave (NSA) y con SA durante un período de seguimiento de 12 meses.

Métodos: El estudio PAGE es un estudio multicéntrico, controlado y observacional realizado en 36 hospitales españoles y dividido en dos fases: una primera fase transversal para la estimación de la prevalencia de AS, y una segunda fase prospectiva (3 visitas en 12 meses) para el seguimiento y caracterización de los pacientes con SA y NSA. Se incluyó un subestudio con ML en 6 hospitales.

Resultados: Se obtuvo una prevalencia de SA del 20,1% entre los pacientes asmáticos, frente al 9,7% de la herramienta ML. La proporción de fenotipos de SA y las características de los pacientes en seguimiento fueron consistentes con estudios anteriores. Las predicciones

clínicas de la evolución de los pacientes fueron poco fiables, mientras que el ML sólo encontró dos modelos predictivos con potencial discriminatorio para predecir resultados.

Conclusión: Este estudio es el primero en estimar la prevalencia del SA, en una población hospitalaria de pacientes con asma, y en predecir los resultados de los pacientes utilizando técnicas estándar y de ML.

Palabras clave: Asma grave. Prevalencia. *Big data*. *Machine learning*. Procesamiento de lenguaje natural. Modelos predictivos.

Summary box

• What do we know about this topic?

Data on the prevalence of severe asthma (SA) are limited. Electronic health records offer a unique research opportunity to test machine learning (ML) tools in epidemiological studies.

• How does this study impact our current understanding and/or clinical management of this topic?

This is the first study to address estimation of the prevalence of SA using both standard and ML techniques. Despite the heterogeneity of our findings, the prevalence of SA in Spanish hospitals using ML techniques may be closer to the actual prevalence of SA in Spain.

Introduction

Asthma remains one of the most common chronic diseases worldwide, and even if a decline in asthma-related hospitalizations and deaths has been reported, the prevalence of the disease continues to increase in many countries [1,2].

Available data on the prevalence of severe asthma (SA) are limited and vary widely between countries [3]. In Spain, the most recent study, from 2011, estimated the prevalence of uncontrolled SA to be 3.9% in adult patients seen in hospital asthma units [4].

Since then, electronic health records (EHRs) have been widely implemented across Spanish hospitals and offer a unique new research opportunity [5,6], since clinical data often appear as structured information. However, analyzing EHRs is usually time-consuming and subject to bias, thus highlighting the suitability of new machine learning (ML) tools (mainly natural language processing [NLP]) for management of this information [7,8].

NLP refers to the branch of artificial intelligence (AI) that aims to make computers able to read and understand text. NLP technologies combine linguistic with statistical and deep learning models to “understand” the full meaning of readable text [9].

Indeed, the use of NLP to extract and analyze the unstructured and structured clinical information in EHRs has helped to advance our clinical and epidemiological understanding of certain diseases [10,11]. However, to the best of our knowledge, few studies have used this technology in patients with SA [12,13].

Therefore, we designed a protocol combining “traditional” methods and ML to assess key, clinically relevant outcomes in SA. The aim of this study was to determine, through chart reviews (manual screening of EHRs by investigators), the proportion of adult asthma patients with SA in outpatient allergy clinics and hospital pulmonology departments in Spain. As secondary objectives, we followed patients up and

described their clinical characteristics over 12 months. We also performed a substudy to incorporate the NLP-based EHR technology Savana to determine the prevalence of SA and predict patients’ clinical course using ML [7,14,15].

Methodology

Design

The “Prevalence of Severe Asthma in Spain” study (PAGE [Spanish initials]) is a multicenter, observational study, split into 2 phases, a cross-sectional phase and a prospective phase, with 2-stage patient selection by random sampling [16]. The research was conducted in 36 hospitals distributed throughout Spain. Patients gave their informed consent to be included in the study, and the protocol was approved by the Ethics Committee of Hospital de La Princesa, Madrid, Spain. The study protocol was registered at ClinicalTrials.gov (ID: NCT03137043). These study findings are reported according to the Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) guideline [17]. The abridged protocol and methodology are available elsewhere [16].

Selection of Participants

For the primary objective, each investigator obtained from their hospital information technology department an internal and confidential list of patients diagnosed with “asthma” (and compatible terms). The contract research organization randomized these lists by hospital. Afterwards, the investigators screened the patients from these lists to verify the asthma diagnosis and the SA diagnosis.

The investigators also recorded sex, age, age at asthma diagnosis, and atopic status of a random sample of these patients (Table 1).

Table 1. Summary of Baseline Demographic Features of the Prev. Severe Asthma Population.

No.	169
Mean (SD) age, y	62.88 (16.87)
Female sex, No. (%)	119 (70.04%)
Mean (SD) age at asthma diagnosis, y	39.53 (18.73)
Respiratory allergy	63 (37.3%)

From the previous randomized lists, the investigators included 12 consecutive patients per site at a ratio of 2:1 for SA vs nonsevere asthma (NSA) (ie, 8 SA patients and 4 NSA patients) (Figure 1).

Asthma and SA were defined according to the GINA guidelines [1]. Both groups of asthma patients were studied and followed up at 3 visits, namely, baseline, 6 months, and 12 months. The information retrieved at baseline and entered into the electronic case report form was the following: demographic and clinical characteristics, including asthma exacerbations, defined as per the American Thoracic Society/European Respiratory Society task force as “the use of systemic corticosteroids, or an increase from a maintenance dose for ≥ 3 days or hospitalization/ER visit because of asthma” [18]; lung function; Asthma Control Test (ACT) score; Saint George’s Respiratory Questionnaire (SGRQ) score; phenotypes according to GEMA guidelines [19]; and

comorbidities, as recorded in the EHRs. Investigators also stated their predictions for the clinical course (change in ACT and SGRQ scores at 6 and 12 months) of patients included in the study at baseline based on their previous clinical experience. These predictions were then individually compared with the patients’ actual clinical course.

Thus, we investigated 3 populations: the “prev. population” (ie, the random lists of patients screened by investigators to determine asthma severity); the “prosp. population” (ie, the population prospectively followed up for 12 months to assess the secondary endpoints), and the “EHRead population” (ie, the EHRs analyzed using EHRead technology, see below).

Substudy With Descriptive and Predictive ML Models

The unstructured clinical information in the EHRs of all patients at the 6 participating sites was extracted and analyzed using EHRead technology.

NLP technology enables the extraction of clinical concepts from EHRs and their subsequent standardization to a common terminology based on Systematized Nomenclature of Medicine – Clinical Terms (SNOMED CT) [20].

Using this information, the period prevalence was estimated, and several ML predictive models were developed to predict clinically relevant events/outcomes in asthma patients, as follows: prescription of add-on biologics, in-hospital mortality, exacerbations, asthma-related visits to the emergency department, and asthma control. The definition of

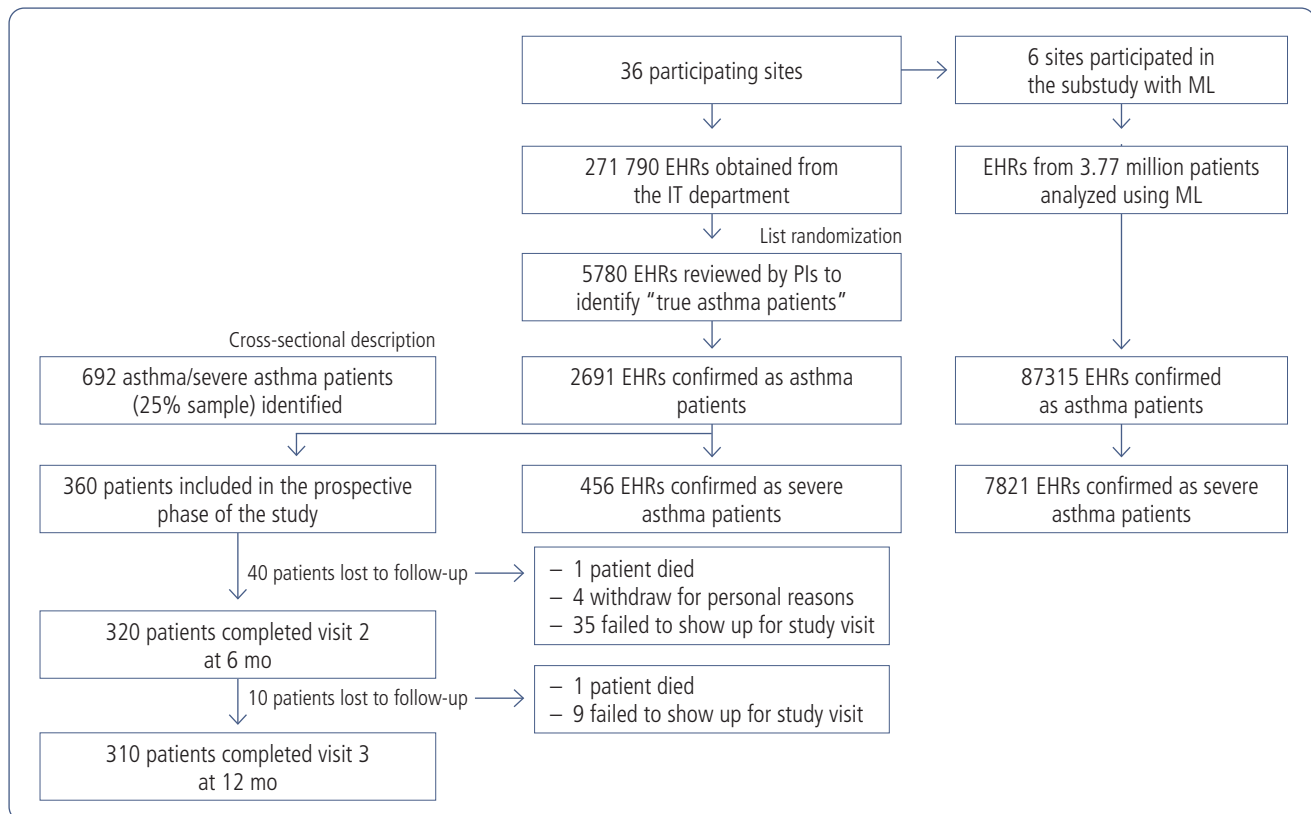


Figure 1. STROBE flowchart of participation in PAGE. EHR indicates electronic health records; ML, machine learning; IT, information technology; PI, principal investigator.

these outcomes is detailed in the Online Appendix. Briefly, the dataset was split into a training dataset (70%) and a test dataset (30%). The features with the highest predictive potential for each of the 6 outcomes were extracted from the training set using random forests. Each of the 6 models was trained on 3 different classification algorithms: multivariable logistic regressions, random forests, and decision tree classifiers. Models were then validated in the test population using metrics such as precision, recall, and F1 score. The best model was chosen based on performance and interpretability. Additional information on the generation of the predictive models is included in the Online Appendix.

Further detail on the selection of participants, setting, variables, sample size calculations, and the statistical analysis are described in the published protocol [16]. Briefly, for the primary objective, a meta-analysis of binary variables (prevalence) was performed according to a fixed-effects model with inverse variance weighting; and for the secondary objectives, univariate and standard bivariate descriptive analyses were performed in the case of categorical or continuous variables, mixed models were run for longitudinal data, and descriptive measures of predictive reliability, probability ratios, and logistic binary multiple regression models were used to detect relevant factors in the predictions of clinical events. The substudy objective was analyzed by determining whether the period-prevalence estimated by the ML was included in 95% of the prevalence provided by the chart reviews.

Table 2. Summary of Baseline Patient Features in the EHRead Severe Asthma Population.

No.	7,821
Mean (SD) age, y	55.5 (19.8)
Female sex, No. (%)	5636 (72.1%)
Smoking status	
Smoker/ex-smoker	2086 (26.6%)
Current active/passive smoker	2457 (30.8%)
Missing	3278 (41.9%)
Comorbidities:	
Chronic rhinitis	194 (2.5%)
Allergic rhinitis	1395 (17.8%)
Anxiety	303 (3.9%)
Depression	1294 (16.5%)
Urticaria	753 (9.6%)
Asthma COPD overlap	1216 (15.5%)
Nasal polyps	766 (9.8%)
Obesity	1451 (18.6%)
Diabetes	1662 (21.3%)
NSAID hypersensitivity	870 (11.1%)
Gastroesophageal reflux Syndrome	902 (11.5%)

Abbreviations: COPD, chronic obstructive pulmonary disease; NSAID, nonsteroidal anti-inflammatory drug.

Results

As shown in Figure 1, a total of 271 790 patients' EHRs showing an asthma-compatible diagnosis were initially obtained by the information technology department. From these, 5780 EHRs were manually screened by the principal investigators to eventually identify 2691 valid asthma patients. This implies a specificity of 46.6% (ie, the proportion of valid asthma patients in the lists obtained from the information technology).

The main features of these populations are shown in Tables 1, 2, and 3 (see Online Appendix for full details). Although no formal comparisons were made between the study populations because they were obtained through different methodologies (ie, EHR screening by principal investigators, patient prospective follow-up, and EHRead), all of them were similar with regards to age distribution and relative sex

Table 3. Summary of Baseline Demographic and Clinical Features of the Prosp. Severe Asthma Population.

No.	231
Mean (SD) age, y	56.9 (15.2)
Female sex, No. (%)	163 (70.6%)
Mean (SD) BMI, kg/m ²	29.4 (6.2)
Smoking status	
Never smoker	155 (67.1%)
Smoker/ex-smoker	76 (32.9%)
Missing	0 (0%)
Mean age at diagnosis of asthma, y	35.3 (17.4)
Family history of asthma	98 (42.4%)
Respiratory allergy	122 (52.8%)
Perennial	93 (76.2%)
Seasonal	29 (23.8%)
Comorbidities:	
None	18 (7.7%)
Atopy	41 (17.6%)
Chronic rhinitis	65 (27.9%)
Allergic rhinitis	74 (31.8%)
Anxiety	37 (15.9%)
Depression	40 (17.2%)
Urticaria	16 (6.9%)
Asthma COPD overlap	9 (3.9%)
Nasal polyps	47 (20.2%)
Obesity	63 (27%)
Diabetes	23 (9.9%)
NSAID hypersensitivity	21 (9%)
Gastroesophageal reflux syndrome	55 (23.6%)

Abbreviations: BMI, body mass index; COPD, chronic obstructive pulmonary disease; NSAID, nonsteroidal anti-inflammatory drug.

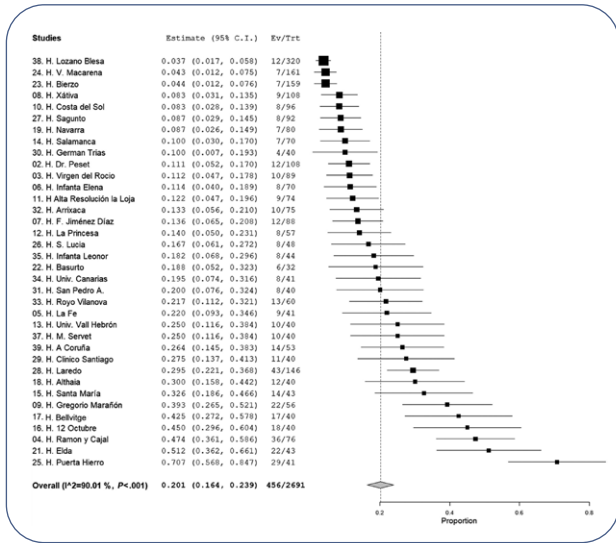


Figure 2. Forest plot of the proportion of severe asthma in all participating sites.

distribution, although they differed in the frequency of certain comorbidities such as allergic rhinitis, diabetes, and anxiety (Table S1 in Online Appendix).

Primary Endpoint

Of the 2691 confirmed adult asthma patients, 456 were confirmed as SA, which results in a global estimated prevalence of SA in Spanish hospitals of 20.1% (95%CI, 0.164-0.239 [range, 3.7%-70.7%]), with high heterogeneity ($I^2 = 89.89\%$) (Figure 2). The high heterogeneity was mostly due to the use of different sampling domains (eg, SA clinics, pulmonology vs allergy outpatient clinics) and encouraged us to perform additional post hoc analyses with the aim of obtaining a nonheterogeneous result (Figures S1 and S2 in the Online Appendix). One such analysis aimed to understand whether the differences between the investigators (eg, allergy vs pulmonology services, coastal vs inland hospitals, large vs small hospitals) were

Table 5. Proportion of Asthma Phenotypes at Baseline.

	SA	NSA	Total
Allergic asthma, No. (%)	97 (43.5%)	72 (58.1%)	169 (48.7%)
Late-onset eosinophilic asthma, No. (%)	60 (26.9%)	16 (12.9%)	76 (21.9%)
Obesity and asthma, No. (%)	31 (13.9%)	13 (10.5%)	44 (12.7%)
Neutrophilic late-onset asthma, No. (%)	22 (9.9%)	8 (6.5%)	30 (8.6%)
Other, No. (%)	13 (5.8%)	15 (12.1%)	28 (8.1%)
Total^a, No. (%)	223 (100%)	124 (100%)	347 (100%)

Abbreviations: NSA, nonsevere asthma; SA, severe asthma.
^aMissing: 10 SA patients and 3 NSA patients. Phenotypes were assigned as per investigator criteria, according to the GEMA 4.1 guideline, which was the current edition at the time the protocol was developed and the data collected (Pearson, $\chi^2=.003$).

significantly influencing the heterogeneity of the result (Table 4). An omnibus *P* value of .189 indicates that none of these factors individually was the cause of the heterogeneity of the primary outcome.

Secondary Endpoints

In the prosp. population, a higher proportion of the allergic phenotype was found in the NSA patients than in the SA patients, while the late-onset eosinophilic phenotype was more frequent in patients with SA (Table 5).

Figure 3 and Table S7 show the change in annualized exacerbation rate, prebronchodilator FEV₁, ACT score, and SGRQ score at 6 and 12 months. As expected, all these clinical endpoints reflected worse disease control in SA patients than in NSA patients. Furthermore, an improvement was observed in both groups, probably owing to closer clinical follow-up of patients during the study and regression towards the mean in patients with SA.

Table 4. Meta-regression of Severe Asthma Prevalence.^a

Covariate	Coefficients	Lower Lim.	Upper Lim.	Std. error	PValue
Intercept	0.190	0.063	0.317	0.065	.003
Allergy	-0.157	-0.274	-0.039	0.060	.009
Pulmonology	0.096	-0.031	0.224	0.065	.138
Emergency department	0.089	-0.043	0.221	0.067	.185
Hospitalization	-0.062	-0.210	0.085	0.075	.408
Coastal	0.061	-0.033	0.155	0.048	.206
Large hospital	0.033	-0.068	0.134	0.052	.523
Omnibus <i>p</i> -Value = 0.189					

^aThe analysis takes into account different variables potentially influencing the result: patient lists (from allergy departments, pulmonology departments, emergency departments only), hospitalizations, coastal vs inland, large vs small hospitals. The omnibus *P* value indicates that none of the factors analyzed has a significant influence on prevalence. Therefore, the observed heterogeneity of the primary endpoint is not attributable to any of these factors.

Table 6. Results of clinical predictions vs actual change in ACT and SGRQ in SA and NSA patients in the prosp. population.^a

SA		ACT change prediction			Total
ACT change		>3 points improvement	No change	>3 points decrease	
>3 points improvement	No. (%)	18 (9.0%)	35 (17.6%)	7 (3.5%)	60 (30.2%)
No change	No. (%)	37 (18.6%)	70 (35.2%)	10 (5.0%)	117 (58.8%)
>3 points decrease	No. (%)	5 (2.5%)	16 (8.0%)	1 (0.5%)	22 (11.1%)
Total	No. (%)	60 (30.2%)	121 (60.8%)	18 (9.0%)	199 (100%)
NSA		ACT change prediction			Total
ACT change		>3 points improvement	No change	>3 points decrease	
>3 points improvement	No. (%)	6 (5.7%)	9 (8.6%)	1 (1.0%)	16 (15.2%)
No change	No. (%)	12 (11.4%)	57 (54.3%)	6 (5.7%)	75 (71.4%)
>3 points decrease	No. (%)	1 (1.0%)	12 (11.4%)	1 (1.0%)	14 (13.3%)
Total	No. (%)	No. (%)	78 (74.3%)	8 (7.6%)	105 (100.0%)
SA		SGRQ change prediction			Total
SGRQ change		>4 points decrease	No change	>4 points increase	
>4 points decrease	No. (%)	12 (6.5%)	76 (41.3%)	25 (13.6%)	113 (61.4%)
No change	No. (%)	3 (1.6%)	38 (20.7%)	4 (2.2%)	45 (24.5%)
No. (%)	No. (%)	3 (1.6%)	21 (11.4%)	2 (1.1%)	26 (14.1%)
Total	No. (%)	18 (9.8%)	135 (73.4%)	31 (16.8%)	184 (100.0%)
NSA		SGRQ change prediction			Total
SGRQ change		>4 points decrease	No change	>4 points increase	
>4 points decrease	No. (%)	2 (2.2%)	34 (37.8%)	6 (6.7%)	42 (46.7%)
No change	No. (%)	1 (1.1%)	31 (34.4%)	2 (2.2%)	34 (37.8%)
>4 points increase	No. (%)	0 (0.0%)	14 (15.6%)	0 (0.0%)	14 (15.6%)
Total	No. (%)	No. (%)	79 (87.8%)	8 (8.9%)	90 (100.0%)

Abbreviations: SA, severe asthma; NSA, nonsevere asthma; ACT, Asthma Control Test; SGRQ, Saint George's Respiratory Questionnaire.

^aThe κ statistic for each of these 4 predictions was 0.021, 0.095, 0.026, and 0.009.

As shown in Table 6, clinicians' predictions on the change in ACT and SGRQ scores in SA and NSA patients at 12 months were unreliable (Cohen κ = 0.021, 0.095, 0.026, and 0.009, respectively).

Investigators were asked on which clinical parameters they based their predictions, and the 3 most common factors mentioned were the number of previous exacerbations, FEV₁, and rescue medication (Online Appendix, Table S4).

Substudy With Descriptive and Predictive ML Models

A total of 3 766 292 EHRs from 6 participating hospitals were analyzed using EHRRead from January 1, 2014 to December 31, 2018. From the total, we identified 87 315 asthma patients, of whom 7821 were diagnosed with SA (population for Table 2).

The period prevalence was measured at the midpoint of the study period (excluding deaths and patients lost to follow-up 1 year or more before the midpoint). A total of 1 681 383 patients visited the study hospitals at least once during this period. Of these, 46 964 had asthma and 4571 had SA, ie, a prevalence of asthma of 2.8%. Among these, the prevalence of SA was 9.7% (Table S2).

Of the 6 predictive models run in the asthma patients identified, only 2 (add-on therapy and in-hospital mortality) showed acceptable discriminatory power to predict outcomes. For these models, performance metrics were slightly lower for outcomes at 12 months than at 6 months in all 3 tested algorithms. For add-on biologics, no significant differences were observed between logistic regression, random forests, and decision tree algorithms, with F1-scores of 0.78 and 0.76 at 6 and 12 months, respectively. For in-hospital mortality, random forests performed best, with F1-scores of 0.81 and

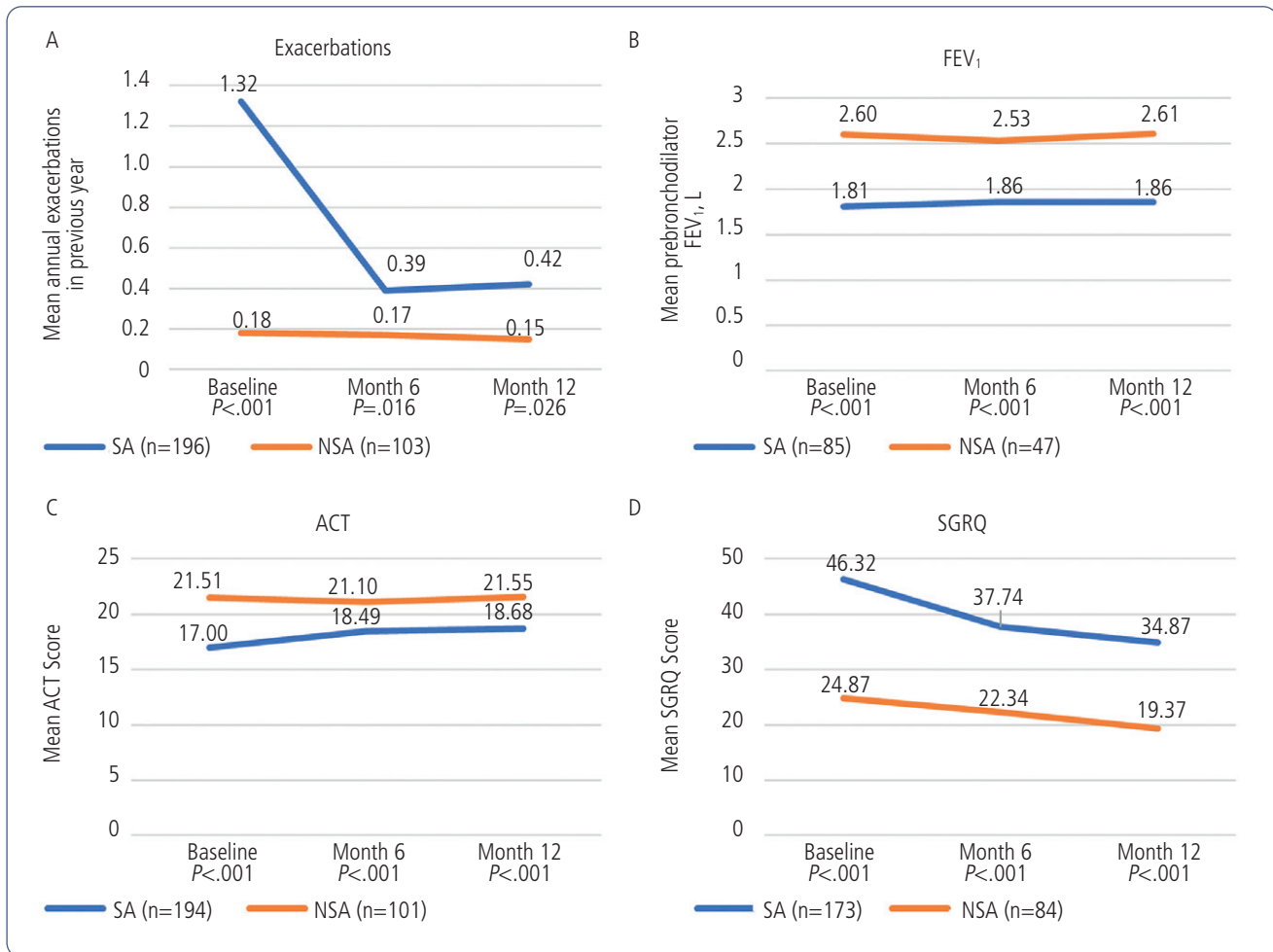


Figure 3. Change in clinical endpoints. All intergroup measures were statistically significant (table S7 in Online Appendix). SA indicates severe asthma; NSA, nonsevere asthma; FEV₁, forced expiratory volume in 1 second; ACT, Asthma Control Test; SGRQ, Saint George's Respiratory Questionnaire.

0.78. However, logistic regression was considered a quasi-equivalent, more interpretable alternative, especially for the 6-month prediction, with an F1-score of 0.8 (see Table S3 for more detail).

For the predictions of add-on therapy with biologics at 6 months, we identified a list of predictor variables that influenced the prediction model. The most relevant factors were atopy, the use of montelukast, and the presence of nasal polyps. Specifically, patients with atopy had an OR of 6.34 (95%CI, 4.21-9.36) for being prescribed an add-on biologic. The OR (95%CI) for montelukast and nasal polyps was 4.59 (3.48-6.10) and 3.97 (2.93-5.35), respectively (Table S5)

For in-hospital mortality, the 3 most relevant factors were being a smoker (OR, 5.48; 95%CI, 1.58-23.96), having a chest x-ray (OR, 3.73; 95%CI, 1.07-16.30), and myocardial infarction (OR, 3.51; 95%CI, 0.80-12.20) (Table S6).

Discussion

PAGE is a clinical study largely based on traditional research methodology that includes in its design a novel

substudy performed using ML, which was applied in parallel to the traditional approach and attempted to estimate the prevalence of SA and to predict patients' clinical course.

The PAGE study showed a prevalence of SA in the hospital setting of 20.1%, which is higher than most previous findings, with significant heterogeneity. The differences found in the prevalence of the study hospitals (range, 3.7%-70.7%) suggest that the main source of heterogeneity arises from the method selected by the information technology departments of the different hospitals, as each region uses its own coding system [21]. These results highlight the need to homogenize clinical practice, data collection, and health coding systems. It is at this point where ML could have shown an advantage because of EHRead, which was performed independently of the hospital platform. Even so, the 2.8% prevalence of asthma estimated by the ML is lower than current estimates [22], while the 9.7% prevalence of SA is higher than most previous publications (Italy, 3.2% [23]; Netherlands, 3.6% [24]; Japan, between 2.4% [25] and 7.8% [26]; Brazil, between 4.1% [27] and 7.6% [28]; Germany, 8.7% [29]; and Sweden, 9.5% [30]). However, most of the studies used different methodologies (eg, hospital records-based studies vs population-based studies)

and definitions for SA, with prevalence ranging from 1.8% to 38% [31], thus hampering comparisons. In an exploratory analysis, we found a reverse correlation between the specificity of EHRs and prevalence (Figure S2), where extrapolating a 100% specificity for EHRs would result in a prevalence of SA of 9.3%.

Analysis of severe asthma phenotypes at baseline showed a higher proportion of the allergic phenotype in the NSA cohort than in the SA cohort, and the opposite was observed for the late-onset eosinophilic phenotype. Along these lines, Pérez de Llano et al [32] found that the most frequent clinical phenotype in an adult Spanish population of patients with uncontrolled SA was late-onset eosinophilic asthma (58.1%).

We observed clinical improvements in exacerbation rate, FEV₁, ACT score, and SGRQ score over the study duration, likely explained by the inclusion of these patients in a study and their subsequent closer clinical follow-up with regression towards the mean. However expected these changes were, they were not reliably predicted by investigators when analyzed individually. We did not find previous studies analyzing the investigators' predictions on the change in their patients' disease course. Research has relied on finding biomarkers to predict disease change. Malinovschi et al [33] used measurement of exhaled fraction of nitric oxide (FeNO) to predict response to inhaled corticosteroids and symptom control in patients with NSA. However, elsewhere, FeNO monitoring was not shown to decrease the frequency of exacerbations or the dose of inhaled corticosteroids in asthma [34]. Castner et al [35] did not rely on physician judgment but instead used fitness and sleep trackers to predict asthma-specific nighttime awakenings and daily FEV₁ changes. The sleep data from the tracker demonstrated predictive ability for daily asthma outcomes.

As for the ML-based prediction about the change in asthma control, we expected to be able to compare the investigators' predictions with monitoring data (gold standard) and the ML predictive model. However, we could not compare the output from the substudy with the standard methodologies, probably because of the lack of predictable patterns in the population, the quality of the medical records analyzed, and the different methods used to build the databases (ie, disease codes vs NLP). The quality of the medical records is a relevant factor that has been worked on thoroughly over the last few decades in other countries. For example, in our substudy with ML, the ACT score was read in only 147 patients (1.9%). This implies that either ACT scores are not commonly stored in EHRs or they are not being interpreted appropriately by ML [36,37]. Our results contrast with other publications describing ML models with large-scale outpatient data that can predict asthma exacerbations [38].

Among the limitations of our study are the heterogeneity of the information and coding systems at the study sites and the fact that it was conducted in a specialized care setting instead of a primary care or population-based setting. Another potential caveat to ML technologies is that their algorithms, data analysis results, and underlying weighting factors sometimes remain opaque (black box methodology in neural networks). Besides, algorithms are also subject to biases resulting from the human use of uncontrolled information (biased samples and labels). Therefore, often, both researchers and algorithms only have access to biased data [39].

Undeniably, ML in particular and computer science in general will enhance the future of research in this area [40]. The main challenge will be to ensure the lack of human bias and heterogeneity in the electronic information that can be analyzed by EHRead technologies.

Conclusion

Ours is the first study to address estimation of the prevalence of SA using both standard and ML techniques. Despite the heterogeneity of our findings, the prevalence of SA in Spanish hospitals was 20.1%, while the approach using ML techniques may be closer to the actual prevalence of severe asthma in Spain, ie, 9.7%, although still higher than in previous studies in other countries.

Acknowledgments

We thank the PAGE Study group members and collaborators who participated in the study and Anna de Prado from IQVIA for her logistic support (contracting management, site training, and database monitoring [funded by GSK]). We thank Savana and Bio-estadística.com for their support in completing this study.

Funding

This study was sponsored by GSK (205807).

Conflicts of Interest

MGSB and DBC are employees of GSK.

CAS participated in speaking activities and advisory boards and provided consultancy services during the period 2015-2019 sponsored by AstraZeneca, Boehringer-Ingelheim, Chiesi, GSK, ALK Mundipharma, Novartis, Pfizer, SEPAR, and NEUMOMADRID. CAS declares not having received, directly or indirectly, funding from the tobacco industry or its affiliates.

JBS participated in speaking activities and advisory boards and provided consultancy services during the period 2015-2020 for Almirall, AstraZeneca, Boehringer-Ingelheim, CHEST, Chiesi, ERS, GEBRO, Grifols, GSK, Linde, Lipopharma, Mundipharma, Novartis, Pfizer, RiRL, Rovi, Sandoz, SEPAR, and Takeda. JBS declares not having received, directly or indirectly, funding from the tobacco industry or its affiliates.

SQ participated in speaking activities and advisory boards and provided consultancy services during the period 2016-2021 sponsored by AstraZeneca, Chiesi, GSK, Mundipharma, Novartis, and Sanofi. SQ declares not having received, directly or indirectly, funding from the tobacco industry or its affiliates.

FAG participated in speaking activities and advisory boards and provided consultancy services during the period 2016-2021 sponsored by AstraZeneca, ALK, Bial, Boehringer-Ingelheim, Chiesi, GSK, Mundipharma, Novartis, Orion-Pharma, and Sanofi. FAG declares not having received, directly or indirectly, funding from the tobacco industry or its affiliates.

CMM and VC declare that they have no conflicts of interest.

References

- Global Initiative for Asthma. Global Strategy for Asthma Management and Prevention. 2021. Available from: www.ginasthma.org.
- GBD Chronic Respiratory Disease Collaborators. Prevalence and attributable health burden of chronic respiratory diseases, 1990-2017: a systematic analysis for the Global Burden of Disease Study 2017. *Lancet Respir Med*. 2020;8:585-96.
- Hassan M, Davies SE, Trethewey SP, Mansur AH. Prevalence and predictors of adherence to controller therapy in adult patients with severe/difficult-to-treat asthma: a systematic review and meta-analysis. *J Asthma*. 2020;57:1379-88.
- Quirce S, Plaza V, Picado C, Vennera M, Casafont J. Prevalence of uncontrolled severe persistent asthma in pneumology and allergy hospital units in Spain. *J Investig Allergol Clin Immunol*. 2011;21:466-71.
- Weegar R. Applying natural language processing to electronic medical records for estimating healthy life expectancy. *The Lancet regional health Western Pacific*. 2021;9:100132.
- Alvarez-Perea A, Sánchez-García S, Muñoz Cano R, Antolín-Amérigo D, Tsilochristou O, Stukus DR. Impact Of "eHealth" in Allergic Diseases and Allergic Patients. *J Investig Allergol Clin Immunol*. 2019;29:94-102.
- Izquierdo JL, Almonacid C, González Y, Del Rio-Bermudez C, Ancochea J, Cárdenas R, et al. The impact of COVID-19 on patients with asthma. *Eur Respir J*. 2021;57: 2003142 .
- Ohno-Machado L. Realizing the full potential of electronic health records: the role of natural language processing. *J Am Med Inform Assoc*. 2011;18:539.
- Education IC. Natural Language Processing (NLP). 2020. Available from: <https://www.ibm.com/cloud/learn/natural-language-processing>.
- Haerian K, Varn D, Vaidya S, Ena L, Chase HS, Friedman C. Detection of pharmacovigilance-related adverse events using electronic health records and automated methods. *Clin Pharmacol Ther*. 2012;92:228-34.
- Izquierdo JL, Morena D, González Y, Paredero JM, Pérez B, Graziani D, et al. Clinical Management of COPD in a Real-World Setting. A Big Data Analysis. *Arch Bronconeumol (Engl Ed)*. 2021;57:94-100.
- Juhn Y, Liu H. Artificial intelligence approaches using natural language processing to advance EHR-based clinical research. *J Investig Allergol Clin Immunol*. 2020;145:463-9.
- Del Rio-Bermudez C, Medrano IH, Yebes L, Poveda JL. Towards a symbiotic relationship between big data, artificial intelligence, and hospital pharmacy. *J Pharm Policy Pract*. 2020;13:75.
- Gomollón F, Gisbert JP, Guerra I, Plaza R, Pajares Villarroya R, Moreno Almazán L, et al. Clinical characteristics and prognostic factors for Crohn's disease relapses using natural language processing and machine learning: a pilot study. *Eur J Gastroenterol Hepatol*. 2022;34:389-97.
- Ancochea J, Izquierdo JL, Soriano JB. Evidence of Gender Differences in the Diagnosis and Management of Coronavirus Disease 2019 Patients: An Analysis of Electronic Health Records Using Natural Language Processing and Machine Learning. *J Womens Health (Larchmt)*. 2021;30:393-404.
- Almonacid Sánchez C, Melero Moreno C, Quirce Gancedo S, Sánchez-Herrero MG, Álvarez Gutiérrez FJ, Bañas Conejero D, et al. PAGE Study: Summary of a Study Protocol to Estimate the Prevalence of Severe Asthma in Spain Using Big Data Methods. *J Investig Allergol Clin Immunol*. 2021;31:308-15.
- von Elm E, Altman DG, Egger M, Pocock SJ, Gøtzsche PC, Vandenbroucke JP. The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: guidelines for reporting observational studies. *Lancet*. 2007;370:1453-7.
- Reddel HK, Taylor DR, Bateman ED, Boulet LP, Boushey HA, Busse WW, et al. An official American Thoracic Society/ European Respiratory Society statement: asthma control and exacerbations: standardizing endpoints for clinical asthma trials and clinical practice. *Am J Respir Crit Care Med*. 2009;180:59-99.
- Plaza Moral V. [GEMA(4.0). Guidelines for Asthma Management. *Arch Bronconeumol*. 2015;51:2-54.
- Anke LE, Tello J, Pardo A, Medrano IH, Ureña A, Salcedo I, et al. Savana: A Global Information Extraction and Terminology Expansion Framework in the Medical Domain. *SEPLN*. 2016;57:23-30.
- Gómez de la Cámara A. [Scientific evidence based medicine: myth and reality of variability in clinical practice and its impact on health outcomes]. *Anales del sistema sanitario de Navarra*. 2003;26:11-26.
- Asher MI, García-Marcos L, Pearce NE, Strachan DP. Trends in worldwide asthma prevalence. *Eur Respir J*. 2020;56:2002094.
- Vianello A, Caminati M, Andretta M, Menti AM, Tognella S, Senna G, et al. Prevalence of severe asthma according to the drug regulatory agency perspective: An Italian experience. *World Allergy Organ J*. 2019;12:100032.
- Hekking PW, Wener RR, Amelink M, Zwinderman AH, Bouvy ML, Bel EH. The prevalence of severe refractory asthma. *J Allergy Clin Immunol*. 2015;135:896-902.
- Sato K, Ohno T, Ishii T, Ito C, Kaise T. The Prevalence, Characteristics, and Patient Burden of Severe Asthma Determined by Using a Japan Health Care Claims Database. *Clin Ther*. 2019;41:2239-51.
- Nagase H, Adachi M, Matsunaga K, Yoshida A, Okoba T, Hayashi N, et al. Prevalence, disease burden, and treatment reality of patients with severe, uncontrolled asthma in Japan. *Allergol Int*. 2020;69:53-60.
- Cançado JED, Penha M, Gupta S, Li VW, Julian GS, Moreira ES. Respira project: Humanistic and economic burden of asthma in Brazil. *J Asthma*. 2019;56:244-51.
- Urrutia-Pereira M, Chong-Neto H, Mocellin LP, Ellwood P, Garcia-Marcos L, Simon L, et al. Prevalence of asthma symptoms and associated factors in adolescents and adults in southern Brazil: A Global Asthma Network Phase I study. *World Allergy Organ J*. 2021;14:100529.
- Taube C, Bramlage P, Hofer A, Anderson D. Prevalence of oral corticosteroid use in the German severe asthma population. *ERJ Open Res*. 2019;5:00092.
- Rönnebjerg L, Axelsson M, Kankaanranta H, Backman H, Rådinger M, Lundbäck B, et al. Severe Asthma in a General Population Study: Prevalence and Clinical Characteristics. *J Asthma Allergy*. 2021;14:1105-15.

31. Caminati M, Senna G. Uncontrolled severe asthma: starting from the unmet needs. *Curr Med Res Opin.* 2019;35:175-7.
32. Pérez de Llano L, Martínez-Moragón E, Plaza Moral V, Trisan Alonso A, Sánchez CA, Callejas FJ, et al. Unmet therapeutic goals and potential treatable traits in a population of patients with severe uncontrolled asthma in Spain. ENEAS study. *Respir Med.* 2019;151:49-54.
33. Malinovschi A, Van Muylem A, Michiels S, Michils A. FeNO as a predictor of asthma control improvement after starting inhaled steroid treatment. *Nitric Oxide.* 2014;40:110-6.
34. Pike K, Selby A, Price S, Warner J, Connett G, Legg J, et al. Exhaled nitric oxide monitoring does not reduce exacerbation frequency or inhaled corticosteroid dose in paediatric asthma: a randomised controlled trial. *Clin Respir J.* 2013;7:204-13.
35. Castner J, Jungquist CR, Mammen MJ, Pender JJ, Licata O, Sethi S. Prediction model development of women's daily asthma control using fitness tracker sleep disruption. *Heart Lung.* 2020;49:548-55.
36. Campbell CM, Murphy DR, Taffet GE, Major AB, Ritchie CS, Leff B, et al. Implementing Health Care Quality Measures in Electronic Health Records: A Conceptual Model. *J Am Geriatr Soc.* 2021;69:1079-85.
37. Neves AL, Freise L, Laranjo L, Carter AW, Darzi A, Mayer E. Impact of providing patients access to electronic health records on quality and safety of care: a systematic review and meta-analysis. *BMJ Qual Saf.* 2020;29:1019-32.
38. Zein JG, Wu CP, Attaway AH, Zhang P, Nazha A. Novel Machine Learning Can Predict Acute Asthma Exacerbation. *Chest.* 2021;159:1747-57.
39. Sun W, Nasraoui O, Shafto P. Evolution and impact of bias in human and machine learning algorithm interaction. *PLoS One.* 2020;15:e0235502.
40. Handelman GS, Kok HK, Chandra RV, Razavi AH, Lee MJ, Asadi H. eDoctor: machine learning and the future of medicine. *J Intern Med.* 2018;284:603-19.

■ *Manuscript received June 17, 2022; accepted for publication August 1, 2022.*

■ **Carlos Almonacid Sánchez**

● <https://orcid.org/0000-0002-1689-3347>

Hospital Universitario de Toledo
Avda. del Río Guadiana
45007 Toledo, Spain
E-mail: caralmsan@gmail.com