Understanding severe asthma through small and Big Data in Spanish hospitals - PAGE Study

Melero Moreno C¹,², Almonacid Sánchez C³, Bañas Conejero D⁴, Quirce S⁵, Álvarez Gutiérrez FJ⁶, Cardona V⁷, Sánchez-Herrero MG⁴, Soriano JB⁸, on behalf of the PAGE Study group⁹

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, and CIBER of Respiratory Diseases (CIBERES), Madrid, Spain
⁶Hospital Universitario Virgen del Rocío, Sevilla, Spain
⁷Hospital Universitario Vall d’Hebron, Barcelona, Spain
⁸Hospital Universitario de La Princesa, Facultad de Medicina, Universidad Autónoma de Madrid, CIBER of Respiratory Diseases (CIBERES), Instituto de Salud Carlos III (ISCIII), Madrid, Spain
⁹Collaborators listed in Supplementary Appendix

Corresponding author:
Carlos Almonacid Sánchez
Hospital Universitario de Toledo, Toledo, Spain
E-mail: caralmsan@gmail.com

This article has been accepted for publication and undergone full peer review but has not been through the copyediting, typesetting, pagination and proofreading process, which may lead to differences between this version and the Version of Record. Please cite this article as doi: 10.18176/jiaci.0848
Summary

Background: Data on severe asthma prevalence is limited. The implementation of Electronic Health Records (EHRs) offers a unique research opportunity to test machine (ML) tools in epidemiological studies. The aim was to estimate severe asthma (SA) prevalence amongst the asthmatic patients seen in hospital asthma units, using both ML and traditional research methodologies. Secondary objectives were to describe non-severe asthma (NSA) and SA patients during a follow-up period of 12 months.

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

Results: A SA prevalence of 20.1% was obtained amongst asthma patients in Spanish hospitals, compared with 9.7% prevalence by the ML tool. The proportion of SA phenotypes and the features of followed-up patients were consistent with previous studies. The clinical predictions of patients’ clinical course was unreliable, while the ML only found two predictive models with discriminatory potential to predict outcomes.

Conclusion: This study is the first to estimate SA prevalence, in a hospital population of asthma patients, and to predict patient outcomes using both standard and ML techniques. These 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. Aprendizaje máquina. Procesamiento del lenguaje natural. Modelos predictivos
Introduction

Asthma remains one of the most common chronic diseases worldwide, and even if some countries have seen a decline in hospitalizations and deaths related to asthma, its prevalence is still increasing in many countries.[1,2]

Available data on the prevalence of severe asthma (SA) is limited and highly variable between countries.[3] In Spain, the most recent study is from 2011 and estimated a 3.9% of severe uncontrolled asthma 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] as much of clinical information appears as structured information. However, analysing EHRs is usually time-consuming and subject to biases, making new Machine Learning (ML) tools (mainly Natural Language Processing, NLP) more suitable to analyse this information [7,8].

Natural Language Processing (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 analyse the unstructured and structured clinical information in patients’ EHRs has helped advancing our clinical and epidemiological understanding of certain diseases [10,11]. However, to the best of our knowledge, there are still few studies that 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 investigator screening of EHRs), the proportion of SA adult patients among asthmatics, in outpatient clinics of allergy and pulmonology hospital departments in Spain. As secondary objectives, we followed up and described their clinical characteristics in a follow-up period of 12
months. In addition, a sub-study aimed to incorporate the NLP-based EHRead® technology Savana to determine SA prevalence and predict patients’ clinical course using ML. [7,14,15]

Methodology

Design
The “Prevalence of Severe Asthma in Spain” study (PAGE, from its Spanish initials) is a multicentre, observational study, split in two phases: cross-sectional and prospective, with a 2-stage subject selection by random sampling [16]. The research was conducted in 36 hospitals geographically distributed throughout Spain. The patients’ informed consent was obtained before their inclusion in the study, and the protocol was approved by the Ethics Committee of Hospital de La Princesa (Madrid, Spain). The study protocol was also posted at the repository 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 (IT) service, an internal and confidential list of patients identified with a diagnosis of “asthma” (and compatible terms). The Clinical Research Organization randomised these lists by Hospital. Afterwards, investigators screened the patients from these lists, checking the asthma diagnosis and severe asthma diagnosis.
Investigators also recorded sex, age, age of asthma diagnosis, and atopy status of a random sample of these patients, as shown in Table 1.
From the previous random lists, investigators included 12 consecutive patients per site in a proportion of 2:1 for SA vs. non-severe asthma (NSA), as per Figure 1 (i.e 8 SA patients and 4 NSA patients).
Asthma and severe asthma were defined according to GINA guidelines (1). Both groups of asthma patients were studied in this observational study and followed up in three visits: baseline, 6 months, and 12 months. The following information was retrieved at baseline and introduced in the electronic Case Report Form (eCRF): demographic and clinical characteristics, including asthma exacerbations, defined as per the ATS/ERS 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), Saint George’s Respiratory Questionnaire (SGRQ), phenotypes according to GEMA guidelines [19], and comorbidities, as recorded in their EHRs. Besides, investigators stated their predictions for the clinical course (change in ACT and SGRQ at 6 and 12 months) of patients included in the study at baseline, these predictions were based in their previous clinical experience. These predictions were then individually compared with patients’ actual clinical course.

Thus, there were three populations investigated within this study: the “prev. population” (i.e., the random lists of patients screened by investigators to determine asthma severity); the “prosp. Population” (i.e., the population prospectively followed up for 12 months to answer the secondary endpoints), and the “EHRead population”, (i.e., the EHRs analysed with the EHRead technology, see below).

Sub-study with descriptive and predictive ML models

The unstructured clinical information in the EHRs of all patients, in 6 participating sites, was extracted and analysed using the EHRead technology. This technology makes it possible to analyze all the EHRs of all the patients who have attended these hospitals.

NLP technology allows for 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: prescription of add-on biologics, in-hospital mortality, exacerbations, asthma-related visits to the Emergency Department, and asthma control. The definition of these outcomes is detailed in the Online Appendix (OA). Briefly, the dataset was split into training (70%) and test (30%) datasets. The features with the highest predictive potential for each of the six outcomes were extracted from the training set using Random Forests. Each of the six models was trained on three different classification algorithms: multivariable logistic regressions, random forest, 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 OA.

The protocol of this study has been published. More details 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 done according to a fixed-effects model with inverse variance weighting, and for secondary objectives, univariate and standard bivariate descriptive analyses were performed for categorical or continuous variables, mixed models for longitudinal data, descriptive measures of predictive reliability and probability ratios and logistic binary multiple regression models were used to detect relevant factors in the predictions of clinical events. The substudy objective was analysed determining whether the period-prevalence estimated by the ML was included in the 95% of the prevalence provided by chart reviews.

Results

As shown in Figure 1, a total of 271,790 patients’ EHRs were initially obtained by the IT services with an asthma-compatible diagnosis. From these, a total of 5,780 EHRs were manually screened by the Principal Investigators (PIs) to eventually identify 2,691 valid asthma patients. This implies a specificity of 46.6% (i.e. the proportion of valid asthma patients in the lists obtained by IT).
The main features of these populations are shown in Tables 1, 2, and 3 (full details in OA). Although no formal comparisons were made between the study populations because they were obtained through different methodologies (i.e. EHR screening by PIs, patient prospective follow up and EHRead), all of them were similar with regards to age distribution and sex relative proportion, while differed in the frequency of certain comorbidities such as allergic rhinitis, diabetes or anxiety (Table S1 in OA).

Primary endpoint

Of the 2,691 confirmed adult asthma patients, 456 were confirmed as SA, which results in a global estimate of 20.1% (95%CI: 0.164, 0.239) prevalence of SA in Spanish hospitals (range: 3.7% - 70.7%), with high heterogeneity, $I^2 = 89.89\%$ (Figure 2). The high heterogeneity, mostly due to ascertainment from different sampling domains (i.e. SA clinics, Pulmonology vs Allergy external consultations, etc) encouraged exploring additional post-hoc analyses aiming to obtain a non-heterogeneous result (Figures S1 and S2 in the OA). One of such analyses aimed to understand if the differences among the investigators: allergy vs pulmonology services, coast vs inland hospitals, big vs small hospitals, etc, were significantly influencing the heterogeneity of the result (table 4). An omnibus p-value of 0.189 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 allergic phenotype was found in the NSA patients compared to 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 FEV$_1$, ACT, and SGRQ 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 is observed in both groups, probably due to a closer clinical follow-up of all patients during the study and regression towards the mean in severe patients.
As shown in Table 6, clinicians’ predictions on the change of ACT and SGRQ scores in SA and NSA patients, at 12 months, were unreliable ([Cohen’s Kappa] = 0.021, 0.095, 0.026, and 0.009, respectively).

Investigators were asked what clinical parameters they used to base their predictions, and the three most common factors mentioned were: the number of previous exacerbations, FEV₁, and rescue medication use (OA Table S4).

Sub-study with descriptive and predictive ML models

A total of 3,766,292 EHRs from 6 participating hospitals were analysed using EHRead throughout the study period. The period for extracting the data was bounded between 1st Jan 2014 and 31st Dec 2018. From these, 87,315 asthma patients were identified, of which 7,821 patients 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 one year or more before the midpoint). 1,681,383 patients visited the study hospitals at least once during this period. Of these, 46,964 were detected as asthma patients and 4,571 as SA, which represented a 2.8% asthma prevalence, and amongst these, a 9.7% prevalence of SA (Table S2).

Out of the six predictive models performed among the identified asthma patients, only two (add-on therapy and in-hospital mortality) showed acceptable discriminatory potential to predict outcomes. For these models, performance metrics were slightly lower for outcomes at 12 months vs. 6 months in all three 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-score of 0.81 and 0.78. However, Logistic Regression was considered a quasi-equivalent, more interpretable alternative, especially for 6-months prediction, with an F1-score of 0.8. See Table S3 for more details.
For the predictions of add-on therapy with biologics at 6 months, a list of predictor variables was detected 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 Odds Ratio of 6.34 (95% CI 4.21-9.36) to be prescribed an add-on biologic. OR (95% CI) for montelukast, and nasal polyps were 4.59 (3.48-6.10) and 3.97 (2.93-5.35) respectively (Table S5)

For in-hospital mortality, the three 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 but includes in its design a novel sub-study with ML, which also attempted to estimate SA prevalence and to predict patients’ clinical course, in parallel to the traditional methodology.

The primary objective of the PAGE study showed a prevalence of severe asthma (20.1%) in the hospital setting, which is higher than most previous findings, with significant heterogeneity. The differences found in the prevalence of different hospitals (range from 3.7% to 70.7%) suggest that the main source of heterogeneity arises from the ascertained selection method by IT services in different hospitals, as each region uses its coding system [21]. These results highlight the need to prioritize the homogenization of clinical practice, data collection and health coding systems. It is at this point, where ML could have shown an advantage because of the EHRead, 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 with 8.7% [29] or Sweden with 9.5% [30]. However, most used different methodologies (e.g. hospital records based studies vs population-based studies) and definitions for severe asthma, with prevalence ranging from 1.8% to 38% [31], making comparisons difficult. In an exploratory analysis, we
found a reverse correlation between the specificity of EHRs and prevalence (Figure S2), where extrapolating a 100% EHRs specificity would result in a 9.3% prevalence of severe asthma.

Analysis of severe asthma phenotypes at baseline showed a higher proportion of allergic phenotype in the non-severe asthma cohort compared to the severe cohort, and the opposite was observed for the late-onset eosinophilic phenotype. In this line, Perez de Llano et al found in an adult Spanish population of patients with severe uncontrolled asthma, that the most frequent clinical phenotype was late-onset eosinophilic asthma (58.1%) [32].

We observed clinical improvements in exacerbation rate, FEV1, ACT, and SGRQ 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, such improvements were not reliably predicted by investigators when analysed individually. We did not find previous studies analysing the investigators’ predictions on the change of their patients’ disease course. Research has relied on finding biomarkers to predict disease change. Malinovschi et al used measurement of exhaled fraction of nitric oxide (FeNO) to predict inhaled corticosteroid (ICS) response and symptom control in patients with non-severe asthma. [33]. In other studies though, FeNO monitoring did not help decrease the frequency of exacerbations or the dose of inhaled corticosteroids in asthma [34]. Castner et al did not rely on physician judgment but instead used fitness & sleep tracker to predict asthma-specific night-time awakenings and daily FEV1 changes. The sleep data form the tracker demonstrated predictive ability for daily asthma outcomes [35].

As for the ML prediction about the change of asthma control, we expected to be able to compare the predictions from investigators, with monitorization data (gold standard) and the ML predictive model. However, we could not compare the output from the sub-study with the standard methodologies. This may be due to several reasons, the lack of predictable patterns in the population, the quality of the medical records analysed, and the different methods used to build the databases (i.e. disease codes vs NLP). The quality of the medical records is a relevant
factor that has been worked on thoroughly over the last decades in other countries. For example, in our sub-study with ML, ACT score was read in only 147 patients (1.9%). This implies that either ACT scores are not commonly stored in EHRs, or the correct interpretation of these scores by ML is still far out of hand [36,37]. Our results contrast with other publications with 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 of the different hospitals involved, 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 coming from the human use of uncontrolled information (biased samples and labels). Therefore, both researchers and algorithms often have only access to biased data [39].

Undeniably, machine learning and, generally, computer science represents an enhancer of research future advance. [40] The main challenge will be ensuring the lack of human bias and heterogeneity in the electronic information analysable by EHRead technologies.

**Conclusion**

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

We thank the PAGE Study group members and collaborators that participated in the study and Anna de Prado from IQVIA for her logistic support (in the form of contracting management, sites’ training, database monitoring; funded by GSK). We thank Savana and Bio-estadistica.com for their support in completing this study.

Funding

This study was sponsored by GSK (205807)

Conflicts of interest

MGSH 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

1. Global Strategy for Asthma Management and Prevention. www.ginasthma.org; 2021.
2. Collaborators GBDCRD. Prevalence and attributable health burden of chronic respiratory diseases, 1990-2017: a systematic analysis for the Global Burden of Disease Study 2017. The Lancet Respiratory medicine. 2020;8(6):585-96.
3. 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. The Journal of asthma : official journal of the Association for the Care of Asthma. 2020;57(12):1379-88.
4. Quirce S, Plaza V, Picado C, Vennera M, Casafont J. Prevalence of uncontrolled severe persistent asthma in pneumology and allergy hospital units in Spain. Journal of investigational allergology & clinical immunology. 2011;21(6):466-71.
5. Weegar R. Applying natural language processing to electronic medical records for estimating healthy life expectancy. The Lancet regional health Western Pacific. 2021;9:100132.
6. Alvarez-Perea A, Sánchez-García S, Muñoz Cano R, Antolin-Amérgo D, Tsilochristou O, Stukus DR. Impact Of "eHealth" in Allergic Diseases and Allergic Patients. Journal of investigational allergology & clinical immunology. 2019;29(2):94-102.
7. Izquierdo JL, Almonacid C, González Y, Del Río-Bermúdez C, Ancochea J, Cárdenas R, et al. The impact of COVID-19 on patients with asthma. Eur Respir J. 2021;57(3).
8. Ohno-Machado L. Realizing the full potential of electronic health records: the role of natural language processing. Journal of the American Medical Informatics Association : JAMIA. 2011;18(5):539-.
9. Education IC. Natural Language Processing (NLP) 2020 [Available from: https://www.ibm.com/cloud/learn/natural-language-processing.
10. 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(2):228-34.
11. 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(2):94-100.
12. Juhn Y, Liu H. Artificial intelligence approaches using natural language processing to advance EHR-based clinical research. The Journal of allergy and clinical immunology. 2020;145(2):463-9.
13. Del Río-Bermúdez 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(1):75.
14. Gomollón F, Gisbert JP, Guerra I, Plaza R, Pajares Villarroja 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. 9000.
15. 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(3):393-404.
16. 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. Journal of investigational allergology & clinical immunology. 2021;31(4):308-15.
17. von Elm E, Altman DG, Egger M, Pocock SJ, Gøtzsche PC, Vandebroucke JP. The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: guidelines for reporting observational studies. The Lancet. 2007;370(9596):1453-7.
18. 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(1):59-99.
19. Plaza Moral V. [GEMA(4.0). Guidelines for Asthma Management]. Arch Bronconeumol. 2015;51 Suppl 1:2-54.
20. 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. 2016;57:23-30.
21. 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(1):11-26.
22. Asher MI, García-Marcos L, Pearce NE, Strachan DP. Trends in worldwide asthma prevalence. Eur Respir J. 2020;56(6).
23. 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(4):100032.
24. Hekking PW, Wener RR, Amelink M, Zwijnderman AH, Bouvy ML, Bel EH. The prevalence of severe refractory asthma. J Allergy Clin Immunol. 2015;135(4):896-902.
25. 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(11):2239-51.
26. 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(1):53-60.
27. 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(3):244-51.
28. 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(3):100529.
29. Taube C, Bramlage P, Hofer A, Anderson D. Prevalence of oral corticosteroid use in the German severe asthma population. ERJ Open Res. 2019;5(4).
30. 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(2):175-7.
32. Pérez de Llano L, Martínez-Moragon E, Plaza Morál 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(2):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(5):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(4):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(12):1019-32.
38. Zein JG, Wu CP, Attaway AH, Zhang P, Nazha A. Novel Machine Learning Can Predict Acute Asthma Exacerbation. Chest. 2021;159(5):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(8):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(6):603-19.
### TABLES

Table 1. Summary of demographic features of SA patients of prev. population

|                       | Prev. SA population |
|-----------------------|---------------------|
| N                     | 169                 |
| Age in years, mean (SD) | 62.88 (16.87)       |
| Sex, female, n, (%)   | 119 (70.04%)        |
| Asthma diagnosis age, years, mean (SD) | 39.53 (18.73)       |
| Respiratory allergy   | 63 (37.3%)          |

*Summary of baseline demographic features of SA patients of prev. population (full details in OA).*
Table 2. Summary of SA patient features in EHRead population.

| EHRead SA population |   |
|----------------------|---|
| N                    | 7,821 |
| Age in years, mean (SD) | 55.5 (19.8) |
| Sex, female, n, (%)    | 5,636 (72.1%) |
| Smoking status         |   |
| • Smoker/ex-smoker     | 2,086 (26.6%) |
| • Current active/passive smoker | 2,457 (30.8%) |
| • Missing              | 3,278 (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 (ACO) | 1216 (15.5%) |
| - Nasal polyps         | 766 (9.8%) |
| - Obesity              | 1451 (18.6%) |
| - Diabetes             | 1662 (21.3%) |
| - NSAID Hypersensitivity | 870 (11.1%) |
| - Gastroesophageal Reflux | 902 (11.5%) |

Summary of baseline SA patient features in EHRead population (full details in OA).
Table 3. Summary of demographic and clinical features of SA patients of prosp. Population

| Prosp. SA population                  |   |
|--------------------------------------|---|
| N                                    | 231|
| Age in years, mean (SD)              | 56.9 (15.2) |
| Sex, female, n, (%)                  | 163 (70.6%) |
| BMI, mean (SD) (kg m^2)              | 29.4 (6.2) |
| Smoking status                        |   |
| • Never smoker                       | 155 (67.1%) |
| • Smoker/ex-smoker                   | 76 (32.9%) |
| • Missing                            | 0 (0%) |
| Asthma diagnosis age, years, mean (SD)| 35.3 (17.4) |
| Family asthma history                | 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 (ACO)          | 9 (3.9%) |
| - Nasal polyps                       | 47 (20.2%) |
| - Obesity                            | 63 (27%) |
| - Diabetes                           | 23 (9.9%) |
| - NSAID Hypersensitivity             | 21 (9%) |
| - Gastroesophageal Reflux            | 55 (23.6%) |

Summary of baseline demographic and clinical features
of SA patients of prosp. population (full details in OA).
Table 4. Meta-regression of Severe Asthma prevalence

| Covariate          | Coefficients | Lower Lim. | Upper Lim. | Std. error | p-Value |
|--------------------|--------------|------------|------------|------------|---------|
| Intercept          | 0.190        | 0.063      | 0.317      | 0.065      | 0.003   |
| Allergy            | -0.157       | -0.274     | -0.039     | 0.060      | 0.009   |
| Pulmonology        | 0.096        | -0.031     | 0.224      | 0.065      | 0.138   |
| ER Department      | 0.089        | -0.043     | 0.221      | 0.067      | 0.185   |
| Hospitalisation    | -0.062       | -0.210     | 0.085      | 0.075      | 0.408   |
| Coast              | 0.061        | -0.033     | 0.155      | 0.048      | 0.206   |
| Big Hospital       | 0.033        | -0.068     | 0.134      | 0.052      | 0.523   |

Omnibus p-Value = 0.189

Meta-regression of Severe Asthma prevalence taking into account different variables potentially influencing the result: patients lists coming from allergy services, pulmonology services, ER Departments only, Hospitalizations, Coast vs inland, big vs small hospitals. The omnibus p-value means that none of the factors analyzed has a significant influence on the prevalence. Therefore, the observed heterogeneity of the primary endpoint is not attributable to any of these factors.
Table 5. Proportion of asthma phenotypes at baseline.

| Phenotype                                         | SA         | NSA        | Total      |
|---------------------------------------------------|------------|------------|------------|
| **Allergic asthma, n (%)**                        | 97 (43.5%) | 72 (58.1%) | 169 (48.7%)|
| **Late-onset eosinophilic asthma, n (%)**         | 60 (26.9%) | 16 (12.9%) | 76 (21.9%) |
| **Obesity and asthma, n (%)**                     | 31 (13.9%) | 13 (10.5%) | 44 (12.7%) |
| **Neutrophilic late-onset asthma, n (%)**         | 22 (9.9%)  | 8 (6.5%)   | 30 (8.6%)  |
| **Other, n (%)**                                  | 13 (5.8%)  | 15 (12.1%) | 28 (8.1%)  |
| **Total*, n (%)**                                 | 223 (100%) | 124 (100%) | 347 (100%) |

*Missing: 10 SA patients and 3 NSA patients.

Phenotypes were assigned as per investigator criteria, according to GEMA 4.1 guideline, which was the current edition at the time of the protocol development and data collection.
Table 6. Results of clinical predictions vs actual change.

| ACT change prediction | >3 points improvement | no change | >3 points decrease | Total |
|-----------------------|-----------------------|-----------|-------------------|-------|
| >3 points improvement | N (%)                 | 18 (9.0%) | 35 (17.6%)        | 60 (30.2%) |
| no change             | N (%)                 | 37 (18.6%)| 70 (35.2%)        | 117 (58.8%) |
| >3 points decrease    | N (%)                 | 5 (2.5%)  | 16 (8.0%)         | 22 (11.1%) |
| Total                 | N (%)                 | 60 (30.2%)| 121 (60.8%)       | 199 (100%) |

| ACT change prediction | >3 points improvement | no change | >3 points decrease | Total |
|-----------------------|-----------------------|-----------|-------------------|-------|
| >3 points improvement | N (%)                 | 6 (5.7%)  | 9 (8.6%)          | 16 (15.2%) |
| no change             | N (%)                 | 12 (11.4%)| 57 (54.3%)        | 75 (71.4%) |
| >3 points decrease    | N (%)                 | 1 (1.0%)  | 12 (11.4%)        | 14 (13.3%) |
| Total                 | N (%)                 | 19 (18.1%)| 78 (74.3%)        | 105 (100.0%) |

| SGRQ change prediction | >4 points decrease | No change | >4 points increase | Total |
|------------------------|-------------------|-----------|-------------------|-------|
| >4 points decrease     | N (%)             | 12 (6.5%) | 76 (41.3%)        | 113 (61.4%) |
| No change              | N (%)             | 3 (1.6%)  | 38 (20.7%)        | 45 (24.5%) |
| >4 points increase     | N (%)             | 3 (1.6%)  | 21 (11.4%)        | 26 (14.1%) |
| Total                  | N (%)             | 18 (9.8%) | 135 (73.4%)       | 184 (100.0%) |
Results of clinical predictions vs actual change in ACT and SGRQ in SA and NSA patients among prosp. population.

*Kappa for each of these four predictions were* $K = 0.021$, $K = 0.095$, $K = 0.026$ and $K = 0.009$. *SA: Severe Asthma; NSA: Nonsevere Asthma.*

| SGRQ change | >4 points decrease | No change | >4 points increase | Total |
|-------------|-------------------|-----------|-------------------|-------|
| >4 points decrease | 2 (2.2%) | 34 (37.8%) | 6 (6.7%) | 42 (46.7%) |
| No change | 1 (1.1%) | 31 (34.4%) | 2 (2.2%) | 34 (37.8%) |
| >4 points increase | 0 (0.0%) | 14 (15.6%) | 0 (0.0%) | 14 (15.6%) |
| Total | 3 (3.3%) | 79 (87.8%) | 8 (8.9%) | 90 (100.0%) |
FIGURES

Figure 1. STROBE flowchart of participation in PAGE. EHRs: Electronic Health Records; ML: Machine Learning; IT Information Technology.
**Figure 2.** Forest plot of the proportion of severe asthma in all participating sites.
**Figure 3.** Change of clinical endpoints. All intergroups measures were statistically significant (table S7 in OA). SA: Severe Asthma; NSA: Nonsevere Asthma; FEV1: Forced Expiratory Volume in 1 second; ACT: Asthma Control Test; SGRQ: Saint George’s Respiratory Questionnaire.