SUPPLEMENTARY MATERIAL

Patient characteristics

	Population prospectively followed-up			EHRead population		Prev. population		
	Severe Asthma	Non-Severe Asthma	Total population	Severe asthma	Total population	Total	Severe asthma	Non-Severe Asthma
N	233	127	360	7821	87315	692	169	523
Age in years, mean (SD)	56.9 (15.2)	50.7 (17.4)	54.7 (16.3)	55.5 (19.8)	49.1 (19.9)	53.1 (21.3)	62.9 (17.5)	50.0 (21.9)
IMC, mean (SD) (kg m²)	29.4 (6.2)	27.5 (6.2)	28.7 (6.3)	-	-	-	-	-
Asthma diagnosis age, years, mean (SD)	35.3 (17.4)	34.8 (19.6)	35.1 (18.2)	-	-	34.2 (21.4)	39.5 (23.0)	31.3 (22.1)
Sex, female, n, (%)	163 (70,6%)	84 (66,7%)	247 (69,2%)	5,636 (72.1%)	57306 (65.6%)	486 (70.0%)	119 (70.0%)	367 (70.2%)
Smoking Status:								
- **Never smoker**	155 (67.1%)	78 (61.9%)	233 (65.3%)	1,314 (16.8%)	-	-	-	-
- **Ex-smoker**	69 (29.9%)	31 (24.6%)	100 (28%)	772 (9.8%)	-	-	-	-
- **Current smoker**	7 (3%)	17 (13.5%)	24 (6.7%)	2,414 (30.8%)	-	-	-	-
Family asthma history	98 (42.4%)	51 (40.5%)	149 (41.7%)	-	-	-	-	-
Respiratory allergy	122 (52.8%)	76 (60.3%)	198 (55.5%)	-	-	346 (53.1%)	63 (37.3%)	283 (54.1%)
- **Perennial**	93 (76.2%)	46 (60.5%)	139 (70.2%)	-	-	-	-	-
- **Seasonal**	29 (23.8%)	30 (39.5%)	59 (29.8%)	-	-	-	-	-
Comorbidities:								
- **None**	18 (7.7%)	19 (15%)	37 (10.3%)	-	-	-	-	-
- **Atopy**	41 (17.6%)	24 (18.9%)	65 (18.1%)	-	-	-	-	-
- **Chronic Rhinitis**	65 (27.9%)	16 (12.6%)	81 (22.5%)	194 (2.5%)	-	-	-	-
- **Allergic Rhinitis**	74 (31.8%)	52 (40.9%)	126 (35%)	1395 (17.8%)	-	-	-	-
- **Anxiety**	37 (15.9 %)	14 (11%)	51 (14.2%)	303 (3.9%)	-	-	-	-
- **Depression**	40 (17.2%)	11 (8.7%)	51 (14.2%)	1294 (16.5%)	-	-	-	-
- **Urticaria**	16 (6.9%)	7 (5.5%)	23 (6.4%)	753 (9.6%)	-	-	-	-
- **Asthma COPD overlap (ACO)**	9 (3.9%)	0 (0%)	9 (2.5%)	1216 (15.5%)	-	-	-	-
- **Nasal polyps**	47 (20.2%)	9 (7.1%)	56 (15.6%)	766 (9.8%)	-	-	-	-
- **Obesity**	63 (27%)	20 (15.7%)	83 (23.1%)	1451 (18.6%)	-	-	-	-
- **Diabetes**	23 (9.9%)	8 (6.3%)	31 (8.6%)	1662 (21.3%)	-	-	-	-
- **NSAID Hypersensitivity**	21 (9%)	7 (5.5%)	28 (7.8%)	870 (11.1%)	-	-	-	-
- **Gastroesophageal Reflux Syndrome**	55 (23.6%)	27 (21.3%)	82 (22,8%)	902 (11.5%)	-	-	-	-

Table S1. Patient characteristics of the three study populations- '-': Missing data

doi: 10.18176/jiaci.0848

Descriptive Machine Learning results

To be included in the severe asthma study population, at least one of the following parameters was present:

- "Severe asthma" terms appeared as such in the EHR (N= 1406; 18%).
- Asthma requiring the use of high doses of ICS, and: a LABA, LAMA, anti-leukotriene, or theophylline in the last 12 months (since the first identifiable asthma diagnosis in EHR) (N= 3614; 46.2%).
- Asthma with continuous treatment with systemic glucocorticoids for 50% or more of the previous year ((since the first identifiable asthma diagnosis in EHR). (N= 3943; 50.4%).
- Patients with asthma undergoing asthma therapy with biologics, except patients treated with biologics without ICS-LABA at high doses. (N= 209; 2.7%)

The percentage of patients who met one or more of the conditions mentioned above was as follows: 1 condition (N= 6712; 85.8%); 2 conditions (N= 918; 11.7%); 3 conditions (N= 140; 1.8%) and 4 conditions (N= 51; 0.7%).

9.1.2. Prevalence

The prevalence of asthma and severe asthma are shown in Table S2. 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). The date "midpoint of the study period" was 15th Jun 2016.

Table S2. The estimated prevalence of asthma and severe asthma.

	Calculation	**Prevalence**
Total hospital population*	1,681,343	NA
Patients with asthma (adults)** / Total hospital population	46,964 / 1,681,343	2.8%
Patients with severe asthma (adults)*** / Total hospital population	4,571 / 1,681,343	0.3%
Patients with severe asthma (adults)/ Patients with asthma (adults)	4,571 / 46,964	9.7%

Prevalence was analyzed as period prevalence, measured at the midpoint of the study, excluding deaths and patients lost to follow-up one year or more before the midpoint.

* Patients who visit the study hospitals at least once between 15th Jun 2015 and 15th Jun 2016 (half of the study period) and who do not die during that period.

** Patients with asthma record between 15th Jun 2015 and 15th Jun 2016 (half of the study period).

*** Patients with severe asthma record between 15th Jun 2015 and 15th Jun 2016 (half of the study period).

doi: 10.18176/jiaci.0848

Predictive model pipeline

The predictive pipeline relies on big data analytics and combines advanced statistics and machine learning tools in the deep-learning spectrum. We developed a prediction model using multivariable logistic regressions, random forests, and decision tree classifiers, which provided an equation or criteria, respectively, to predict an individual's risk for a specific event based on their clinical information.

The pipeline used for the generation of the predictive models presented here includes the following steps:

1. Creation of a table containing a subset of the aggregated database. This table represents a subset of the aggregated database produced via the NLP pipeline and only includes data from the severe asthma population. In this table, each row represents a single (anonymous) patient, and columns contain the variables included in the model. This table contains all the data that was processed to generate the model.
2. Optimization of the table containing a subset of the aggregated database. Because not all variables were available for each patient included in the study, missing values in the table were processed and dealt with before any algorithm was implemented. In this process, we distinguished three types of variables, namely binary variables (yes/no presence of the variable or term), numeric variables (i.e., laboratory values), and multi-class variables (i.e., asthma control: good, bad, or regular).
    a) Missing values. For numeric variables, all variables with a relative percentage of missing data (i.e., missing observations for that variable) > 50% were removed from further analyses and therefore not included in the predictive model. This step was aimed at eliminating statistical noise when predicting the desired outcome. Variables excluded encompass the following:
        - FEV1 preBD (%)
        - FEV1/FVC preBD (%)
        - FVC preBD (%)
        - FeNO (fractionated exhaled nitric oxide level)
        - Eosinophils
        - BMI
        - Weight
        - Neutrophils
        - FEV1 preBD (cc)
        - FVC preBD (cc)
        - Total IgE
        - FEV1 postBD (%)
        - FVC postBD (%)
    b) Imputation of missing data. The procedures used for the imputation of missing data vary across variable types:
        - For binary variables, missing data were treated as true '0' values (i.e., no apparition of the variable)
        - For numeric variables, missing data were filled with the median value for existing data
        - For multi-class variables, missing data were filled with the median value for existing data

doi: 10.18176/jiaci.0848

c) Definition of dependent and independent variables. In this step, the variables that the model aimed to predict (i.e., dependent variables) were removed from the list of potential predictors (i.e., independent variables).
d) Feature selection algorithm (dimensionality reduction). To guarantee that the model's output was easily interpretable from a clinical standpoint, the optimal number of independent variables (or features) must be between 10 and 20. Several feature extraction algorithms were used to obtain the right amount of features, including
    - Random forest used to extract the top 10-20 variables to be included in the prediction
    - Recursive Feature Elimination algorithms, based on decision trees or logistic regression
    - Mutual information algorithms.
e) Prediction algorithms. The final stage in the generation of the algorithm involved the following procedures:
    - Split of the data. To train and validate the models, the dataset was separated into a 70/30 training/validation split. Meaning that 70% of the data was used to train and fine-tune the weights of the variables and the remaining 30% was used to validate or test the model's performance.
    - Data augmentation. Upsampling or downsampling techniques were used to balance the number of positive and negative cases for each given outcome. This step was performed before training the model.
    - Output. Three types of models were trained based on their clinical interpretability: random forest, decision trees and logistic regression.

Finally, the predictive models were assessed in terms of F1-value, precision, recall, and accuracy.

$$F1 - score = \frac{2XPrecisionXRecall}{Precision + Recall}$$

EVENT TO PREDICT	Models' performance					
	Logistic Regression		Random Forest		Decision Tree	
	6 M	12 M	6 M	12 M	6 M	12 M
Add-on biologics	**0.78**	**0.76**	**0.78**	**0.76**	**0.78**	**0.75**
In-hospital mortality	**0.80**	**0.75**	**0.81**	**0.78**	**0.77**	**0.71**
Exacerbations	0.54	0.54	0.57	0.51	0.54	0.47
Change of treatment	0.51	0.53	0.50	0.51	0.51	0.57
Asthma-related visits to ER	0.55	0.53	0.57	0.57	0.48	0.5
Asthma control	0.39	0.4	0.39	0.4	0.39	0.4

Table S3. F-1 Score of Savana predictive models on the study population. A robust score is considered when the performance is >0.7.

doi: 10.18176/jiaci.0848

Determinant factors to support clinicians' predictions

Percentage of investigators that marked determinant factors as supporting factors for their predictions			
	Severe Asthma	Non-Severe Asthma	**Total population**
FEV1	87.6%	85.0%	86.7%
Biologic treatment addition	33.5%	26.8%	31.1%
Eosinophils (blood or sputum)	47.2%	41.7%	45.3%
%FEV1 reversibility	59.7%	59.1%	59.4%
Concomitant diseases	44.6%	39.4%	42.8%
ACT	68.2%	66.9%	67.8%
Exacerbation number	91.8%	84.3%	89.2%
Comorbidities	67.4%	59.8%	64.7%
Hospitalization number	72.5%	67.7%	70.8%
Rescue medication use	82.4%	82.7%	82.5%
Asthma symptoms	66.1%	65.4%	65.8%
Emergency visits	76.4%	75.6%	76.1%
Treatment changes	53.6%	58.3%	55.3%
Tobacco use	39.1%	42.5%	40.3%
Mortality	16.7%	14.2%	15.8%
ICS dose	57.1%	58.3%	57.5%
Need of OCS	74.7%	64.6%	71.1%
Inhaler technique	41.6%	45.7%	43.1%
FeNO	60.9%	54.3%	58.6%
BMI	33.5%	32.3%	33.1%
COPD	20.2%	21.3%	20.6%
Adherence to treatment	61.8%	60.6%	61.4%
GINA therapeutic step	40.3%	40.2%	40.3%
Sex	6.9%	7.1%	6.9%
Race	0.4%	0.0%	0.3%
Education level	13.7%	10.2%	12.5%
Rural lifestyle	10.7%	11.0%	10.8%

Table S4. Factors that support clinicians' predictions.

These were the features that investigators marked as relevant when establishing their predictions on patient’s future clinical course.

doi: 10.18176/jiaci.0848

The univariate Odds ratio for add-on biologics at 6 months

Variable	*Estimate (95% CI)*	*p-value*
Atopy	6.34 (4.21, 9.36)	0.000**
Montelukast	4.59 (3.48, 6.10)	0.000**
Nasal polyps	3.97 (2.93, 5.35)	0.000**
Tiotropium bromide	2.95 (2.21, 3.91)	0.000**
Theophylline	4.11 (2.47, 6.57)	0.000**
Chest CT	1.9 (1.32, 2.69)	0.000**
Systemic glucocorticoids	1.64 (1.23, 2.21)	0.001**
Formoterol	0.71 (0.32, 1.39)	0.372
Sinusitis	3.28 (2.35, 4.53)	0.000**
Fluticasone	3.79 (2.56, 5.50)	0.000**
Atrial fibrilation	0.21 (0.07, 0.51)	0.000**
Arterial hypertension	0.41 (0.30, 0.55)	0.000**
Heart failure	0.37 (0.20, 0.66)	0.000**

*Table S5. Variables that influence the prediction of whether a patient will be prescribed an add-on therapy with biologic therapies in the next 6 months. Positive variables increase likelihood of event whereas negative variables decrease it. **Coefficients must not be interpreted directly: for the calculation of probabilities, a logistic function must be applied to the result of the model for proper interpretations.*

The univariate Odds ratio for Mortality at 6 months

Variable	*Estimate (95% CI)*	*p-value*
Depression	2.88 (0.76, 9.59)	0.062
Diabetes	2.04 (0.54, 6.78)	0.197
Systemic glucocorticoids	1.92 (0.55, 8.41)	0.294
Myocardial infarction	3.51 (0.80, 12.20)	0.048**
Formoterol	1.47 (0.03, 9.86)	0.511
Ipratropium bromide	2.69 (0.77, 8.86)	0.096
Chest X-Ray	3.73 (1.07, 16.30)	0.026**
Chest CT	2.30 (0.41, 8.72)	0.181
Asthma with obesity	3.31 (0.95,10.91)	0.031**
Smoker	5.48 (1.58, 23.96)	0.002**
Heart failure	1.86 (0.33, 7.06)	0.409
Malignant neoplasms	2.72 (0.71, 9.04)	0.074
Palpitations	2.30 (0.41, 8.75)	0.179

*Table S6. Variables that influence the prediction of in-hospital death in the next 6 months. Positive variables increase the likelihood of an event, whereas negative variables decrease it. **Coefficients must not be interpreted directly: for the calculation of probabilities, a logistic function must be applied to the result of the model for proper interpretations.*

doi: 10.18176/jiaci.0848

Prevalence analysis by regions.

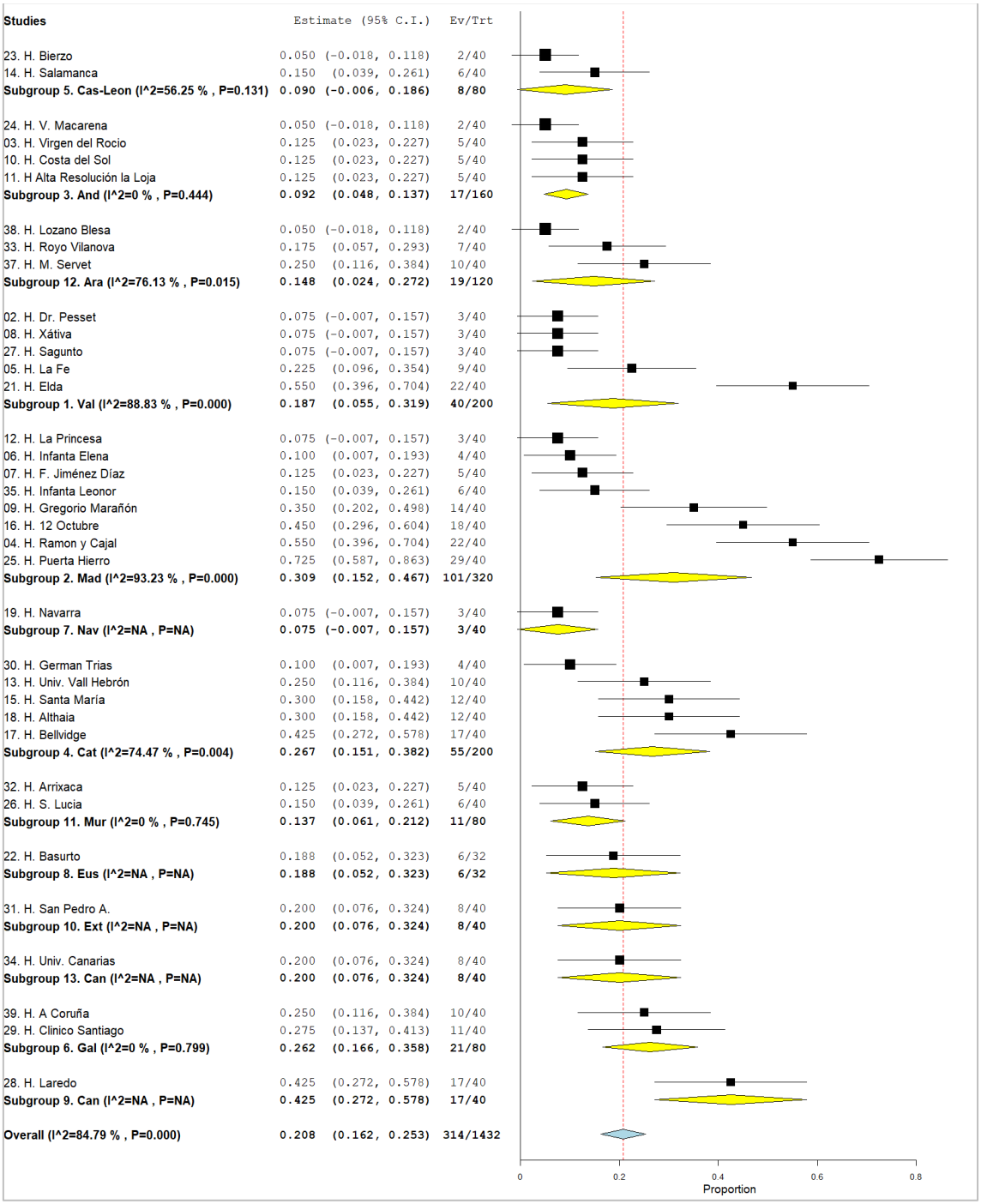

Figure S1. Forest plot of severe asthma prevalence. Subgroup analysis by regions

doi: 10.18176/jiaci.0848

Analysis of the influence of specificity on the Prevalence

Covariate	Coefficients	Low Lim.	Upp Lim.	Std.error	p-Value
Intercept	0.348	0.245	0.451	0.052	< 0.001
Especificity	-0.255	-0.423	-0.088	0.086	0.003

Omnibus p-Value
0.003

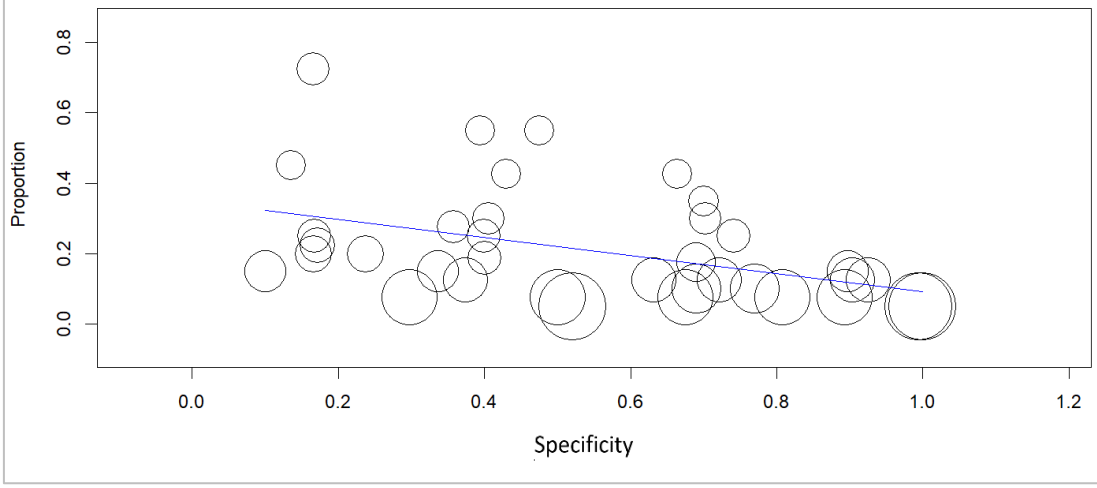

Figure S2. Analysis of the influence of Specificity on the Prevalence.

Specificity (the proportion of true asthma patients in the lists ascertained by IT services) was found to have significant influence on the prevalence. Higher specificity (higher quality of the site's eCRs classification) meant lower prevalence. Interpolating results, a 100% specificity would result in a 9,3% prevalence of Severe Asthma. However, this is a theoretical value obtained by extrapolation of the model, adjusted by Specificity.

doi: 10.18176/jiaci.0848

Clinical follow-up of prosp. Population

		Baseline			6 months			12 months		
		SA	NSA	P value	SA	NSA	P value	SA	NSA	P value
Annualized exacerbation rate (n=196/103)*	Mean (SD)	1.32 (2.3)	0.18 (0.5)	p<0.001	0.39 (0.8)	0.17 (0.5)	p=0.016	0.42 (1.1)	0.15 (0.7)	p=0.026
Pre-BD FEV_1 (n=85/47)*	Mean (SD), L	1.81 (0.7)	2.60 (0.8)	p<0.001	1.86 (0.8)	2.53 (0.8)	p<0.001	1.86 (0.8)	2.61 (0.8)	P<0.001
ACT score (n=194/101)*	Mean (SD)	17.00 (5.6)	21.51 (3.8)	p<0.001	18.49 (5.1)	21.10 (3.8)	p<0.001	18.68 (5.3)	21.55 (4.0)	p<0.001
SGRQ score (n=173/84)*	Mean (SD)	46.32 (21.7)	24.87 (16.4)	p<0.001	37.74 (22.0)	22.34 (16.8)	p<0.001	34.87 (23.0)	19.37 (16.2)	p<0.001

Table S7. Change of clinical endpoints in patients prospectively followed-up during study period.

doi: 10.18176/jiaci.0848

List of Study Investigators

- **Hospital Dr. Peset:** Eva Martínez Moragón
- **Hospital Virgen del Rocío:** Francisco Javier Álvarez Gutiérrez, Juan Francisco Medina Gallardo, Auxiliadora Romero, Krasimira Baynova, Maria Victoria Maestre
- **Hospital Ramón y Cajal:** Carlos Almonacid Sánchez
- **Hospital La Fe:** Miguel Ángel Díaz Palacios
- **Hospital Infanta Elena:** Aythamy Henríquez Santana
- **Hospital Fundación Jiménez Díaz:** Mª Mar Fernández Nieto
- **Hospital de Xátiva: Luis Ángel Navarro Seisdedos**
- **Hospital Gregorio Marañón:** Luis Puente, Wather Girón Matute
- **Hospital Costa del Sol:** Alicia Padilla Galo
- **Hospital Alta Resolución Loja:** Bernardino Alcázar
- **Hospital La Princesa:** Carolina Cisneros
- **Hospital Vall d'Hebron:** Victoria Cardona Dahl, Olga Luengo
- **Hospital 12 de Octubre**: Rocío Díaz Campos, Carlos Melero Moreno
- **Hospital A Coruña:** Marina Blanco Aparicio
- **Hospital La Paz:** Santiago Quirce Gancedo,
- **Hospital Puerta de Hierro:** Antolín López Viña, Andrea Trisán Alonso, Teresa Caruana Careaga,
- **Hospital Virgen Macarena**: Mª Carmen Segura
- **Hospital Salamanca**: Ignacio Dávila
- **Hospital Santa María de Lleida**: Lluís Marques Amat
- **Hospital Bellvitge**: Ramón Lleonart Bellfill
- **Hospital Althaia Manresa:** María Peña Peloche
- **Complejo Hospitalario de Navarra:** Pilar Cebollero Rivas
- **Hospital de Elda:** Ana Isabel Gutierrez Rubio
- **Hospital de Basurto:** Aizea Mardones
- **Hospital de El Bierzo:** Juan Ortiz de Saracho
- **Hospital Santa Lucía:** Francisco Javier Bravo
- **Hospital Sagunto:** Marta Palop
- **Hospital de Laredo:** Juan Luis García Rivero
- **Hospital Clínico de Santiago:** Francisco Javier González-Barcala
- **Hospital Germans Trias i Pujol:** Carlos Martínez Rivera
- **Hospital San Pedro de Alcántara:** Agustín Sojo González
- **Hospital de Arrixaca:** Jose Damián López Sánchez
- **Hospital Royo Vilanova:** Jose Ángel Carretero García
- **Hospital Universitario de Canarias:** Paloma Poza Guedes, Ruperto González Pérez
- **Hospital Infanta Leonor:** Beatriz Arias Arcos
- **Hospital Miguel Servet:** Elisabet Vera Solsona
- **Hospital Lozano Blesa:** Carlos Colás

doi: 10.18176/jiaci.0848