Original Article

Cluster Analyses From the Real-World NOVELTY Study: Six Clusters Across the Asthma-COPD Spectrum

Rod Hughes, MD[a], Eleni Rapsomaniki, PhD[b], Aruna T. Bansal, PhD[c], Jørgen Vestbo, MD[d], David Price, FRCGP[e,f], Alvar Agustí, MD[g], Richard Beasley, DSc[h], Malin Fageras, PhD[i], Marianna Alacqua, PhD[b], Alberto Papi, MD[j], Hana Müllerová, PhD[b], and Helen K. Reddel, PhD[k]; for the NOVELTY Scientific Community and the NOVELTY study investigators* *Cambridge and Aberdeen, United Kingdom; Singapore; Barcelona, Spain; Wellington, New Zealand; Gothenburg, Sweden; Ferrara, Italy; and Sydney, NSW, Australia*

What is already known about this topic? Asthma and chronic obstructive pulmonary disease (COPD) are heterogeneous, and many patients have features of both conditions, but the treatment recommendations are markedly different. For precision medicine, we need to identify distinct disease phenotypes or underlying molecular mechanisms.

What does this article add to our knowledge? Cluster analysis among patients with asthma and/or COPD in the large, real-world NOVEL observational longiTudinal studY yielded 6 overlapping clusters, each including asthma, asthma+COPD, and COPD, with several discriminatory features that differed from conventional diagnostic characteristics.

How does this study impact current management guidelines? The findings demonstrate the heterogeneity and overlap between features of asthma, asthma+COPD, and COPD in real-world patients, contrasting with the siloed populations with asthma or COPD on which current management guidelines are based.

BACKGROUND: Asthma and chronic obstructive pulmonary disease (COPD) are complex diseases, the definitions of which overlap.
OBJECTIVE: To investigate clustering of clinical/physiological features and readily available biomarkers in patients with physician-assigned diagnoses of asthma and/or COPD in the NOVEL observational longiTudinal studY (NOVELTY; NCT02760329).
METHODS: Two approaches were taken to variable selection using baseline data: approach A was data-driven, hypothesis-free and used the Pearson dissimilarity matrix; approach B used an unsupervised Random Forest guided by clinical input. Cluster analyses were conducted across 100 random resamples using partitioning around medoids, followed by consensus clustering.
RESULTS: Approach A included 3796 individuals (mean age, 59.5 years; 54% female); approach B included 2934 patients (mean age, 60.7 years; 53% female). Each identified 6 mathematically stable clusters, which had overlapping characteristics. Overall, 67% to 75% of patients with asthma were in 3 clusters, and approximately 90% of patients with COPD were in 3 clusters. Although traditional features such as allergies and current/ex-smoking (respectively) were higher in these clusters, there were differences between clusters and approaches in features such as sex, ethnicity, breathlessness,

[a]Early Clinical Development, AstraZeneca, Cambridge, United Kingdom
[b]BioPharmaceuticals Medical, AstraZeneca, Cambridge, United Kingdom
[c]Acclarogen, Cambridge, United Kingdom
[d]University of Manchester, Manchester, United Kingdom
[e]Observational and Pragmatic Research Institute, Singapore
[f]Centre of Academic Primary Care, Division of Applied Health Sciences, University of Aberdeen, Aberdeen, United Kingdom
[g]Respiratory Institute, Hospital Clinic, University of Barcelona, IDIBAPS, CIBERES, Barcelona, Spain
[h]Medical Research Institute of New Zealand, Wellington, New Zealand
[i]BioPharmaceuticals Medical, AstraZeneca, Gothenburg, Sweden
[j]Respiratory Medicine Unit, Department of Translational Medicine, Università di Ferrara, Ferrara, Italy
[k]The Woolcock Institute of Medical Research and the University of Sydney, Sydney, Australia
*The names of the NOVELTY Scientific Community and the NOVELTY study investigators are provided in this article's Online Repository at www.jaci-inpractice.org.

```
Abbreviations used
    BMI- body mass index
    COPD- chronic obstructive pulmonary disease
    Feno- fractional exhaled nitric oxide
    FEV₁- forced expiratory volume in 1 second
NOVELTY- NOVEL observational longiTudinal studY
```

frequent productive cough, and blood cell counts. The strongest predictors of the approach A cluster membership were age, weight, childhood onset, prebronchodilator FEV$_1$, duration of dust/fume exposure, and number of daily medications.

CONCLUSIONS: Cluster analyses in patients from NOVELTY with asthma and/or COPD yielded identifiable clusters, with several discriminatory features that differed from conventional diagnostic characteristics. The overlap between clusters suggests that they do not reflect discrete underlying mechanisms and points to the need for identification of molecular endotypes and potential treatment targets across asthma and/or COPD.

INTRODUCTION

Asthma and chronic obstructive pulmonary disease (COPD) are the 2 most prevalent obstructive lung diseases, contributing a substantial burden to patients and to health care systems globally.[1] They have long been recognized as heterogeneous and overlapping conditions[2] and are now thought to comprise a spectrum of lung diseases with multiple poorly defined clinical phenotypes.[3-5] Between 4% and 66% of patients with obstructive lung disease have features associated with both asthma and COPD or have been given both diagnoses.[6,7] Treatment decisions for clinicians are therefore difficult, because of the opposite recommendations in asthma and COPD guidelines about safety of treatment with long-acting bronchodilators alone, on the basis of randomized controlled trials conducted in the highly selected, nonoverlapping populations of asthma *or* COPD currently required by regulators.[8-11] Most research investigating underlying mechanisms is also restricted to patients with a diagnosis of asthma *or* COPD. However, to provide a precision-medicine, biology-led approach to treatment,[12] it is important to identify distinct disease subtypes (phenotypes) or precise underlying molecular mechanisms that are associated with distinct outcomes or treatment responses (endotypes).

Clustering analyses are hypothesis-generating methods that typically apply complex mathematical modeling to group patients by a range of variables. Clustering methodology varies, particularly with regard to the process of selecting the clinical variables for inclusion in the analysis. Variable selection may be "facilitated" by expert input or "nonfacilitated" and data-driven.[13,14] The overall aim of clustering is to minimize variance between patients included in each group while maximizing differences between groups.[15]

The NOVEL observational longiTudinal studY (NOVELTY) is a large, global, prospective, observational study of patients from primary or specialist care with a physician-assigned diagnosis or a suspected diagnosis of asthma and/or COPD.[16,17] The primary objectives of NOVELTY are to describe patient characteristics, treatment patterns, and burden of illness and to identify the clinical phenotypes and molecular endotypes (on the basis of biomarkers and/or clinical parameters) that are associated with differential outcomes for symptom burden, clinical evolution, and health care utilization over time. Here, as an initial step, we

conducted prespecified and exploratory cross-sectional cluster analyses, using nonfacilitated and facilitated approaches to variable selection, in patients with physician-assigned diagnoses of asthma and/or COPD in NOVELTY. The aim was to investigate whether these clusters would identify distinct patient groups that might indicate specific underlying pathways or suggest future treatment targets, independent of conventional diagnostic labels.

METHODS

Study design

The NOVELTY study design has been reported previously.[16] NOVELTY (NCT02760329) is a global, prospective, observational 3-year study that enrolled more than 11,000 patients 12 years or older with a physician-assigned diagnosis or a suspected diagnosis of asthma, COPD, or both (asthma+COPD) from primary and specialist clinical practices in 19 countries. Patients were excluded only if their primary respiratory diagnosis was not asthma or COPD, if they had participated in a respiratory interventional trial during the previous 12 months, or if, in the physician's opinion, they were unlikely to complete 3 years of follow-up. To avoid the high selectivity of regulatory clinical trials[8-11] and allow generalizability to patients in clinical practice, the study included patients treated for asthma and/or COPD in primary or specialist care, without specifying any conventional diagnostic or severity criteria. To ensure sufficient patients with severe disease, enrollment was stratified by physician-assigned diagnosis and physician-assessed severity (mild, moderate, or severe). The study protocol was approved in each country by the relevant institutional review board, and all patients provided written informed consent.

Baseline NOVELTY data set

The data set for these cluster analyses comprised the baseline clinical, functional, and initial biomarker data for patients enrolled as of March 5, 2018, from 18 countries (data from China were excluded because of a change in data transfer regulations in May 2019). The data set included more than 170 demographic, clinical, and biomarker variables, but physician-assigned diagnosis, physician-assessed severity, and medications were excluded from the cluster analysis to avoid bias from preexisting clinical concepts. Details of the methods for selection of variables and clustering are summarized herein, with full details in this article's Online Repository at www.jaci-inpractice.org.

Input variable selection

Two approaches were taken to variable selection (see Table E1 in this article's Online Repository at www.jaci-inpractice.org). The first (approach A) was a nonfacilitated, hypothesis-free approach[13,14]; the framework is shown in Figure E1, A, in this article's Online Repository at www.jaci-inpractice.org. All clinical variables were plotted using histograms and a list of mean-shift outliers was derived using studentized residuals at a Bonferroni-adjusted P value of less than .001. After excluding extreme outliers, the following criteria were applied to the variables in the baseline data set: less than 10% missing data; data either continuously distributed or, if binary, with minor frequency greater than 5%. All clinical variables excluded because of missing data were confirmed to be represented by a significantly correlated variable that was more complete. A total of 79 clinical variables satisfied these criteria. Simulated annealing and genetic algorithm methods were then applied to arrive at a low-correlation variable set to act as surrogate for the whole.[18] The final set of 13 input variables for cluster analysis included the

following: age at onset of respiratory symptoms (years), basophil count, body mass index (BMI), dyspnea (modified Medical Research Council dyspnea grade), fractional exhaled nitric oxide (F_{ENO}), hemoglobin, hospital admission for exacerbation in the last 12 months, number of exacerbations in the last 12 months, prebronchodilator FEV_1, smoking status, use of antibiotics for exacerbation in the last 12 months, white blood cell count, and years of education. Figure E2 (in this article's Online Repository at www.jaci-inpractice.org) shows Kendall's τ correlations within the data set tested after selection of these variables. Eigenvectors were derived to ensure orthogonality, and the Pearson dissimilarity matrix was input into the cluster analysis.

In the second approach (approach B), selection of variables for clustering was guided by clinical relevance and incorporated a nonparametric approach to produce a dissimilarity matrix (unsupervised Random Forest). The framework is shown in Figure E1, B. Variables were prefiltered to a nonredundant set of candidate variables considered relevant to the analysis aims. The resulting 65 variables then underwent variable clustering analysis for mixed data, on the basis of principal- component analysis for continuous variables and factor analysis for nominal variables, using R package clustOfVar[19]; Figure E3 (in this article's Online Repository at www.jaci-inpractice.org) shows the cluster variable dendrogram. On the basis of clinical expert input (authors R.H. and H.M.), 1 variable was selected from each of the 27 variable clusters. The final set of 27 input variables for cluster analysis included the following: age at baseline (years), biomass fuels ever used for cooking/heating (including open fire burning wood, charcoal, coal, or other biomass fuel), blood eosinophil count ($\times 10^9$/L), blood neutrophil count ($\times 10^9$/L), BMI, chronic bronchitis, coronary heart disease or heart failure, current smoker, depression or anxiety or antidepressant use, exacerbations in the last 12 months (n), family history (asthma/COPD or allergies), F_{ENO} (parts per billion), hospital admissions for exacerbations in the last 12 months (n), hospital visits (emergency department or admissions) for other reasons in the last 12 months (n), lymphocyte count ($\times 10^{12}$/L), nonallergic rhinitis/sinusitis, number of daily maintenance medications for asthma/COPD in the last 12 months, parental smoking (father only, mother only, both, neither/unknown), prebronchodilator FEV_1 (% predicted), prebronchodilator forced vital capacity (% predicted), respiratory allergies (including allergic rhinitis, seasonal/perennial rhinitis/sinusitis, eye, or mold allergies), sex, St George's Respiratory Questionnaire activity score, St George's Respiratory Questionnaire symptoms score, smoking pack-years, time (years) since respiratory symptom onset or diagnosis (whichever was earlier), and visits to specialist in the last 12 months (n). Figure E4 (this article's Online Repository at www.jaci-inpractice.org) shows Kendall's τ correlations within the data set tested after selection of these variables. These variables were included in a Random Forest dissimilarity matrix, which was input into the cluster analysis.

Clustering

Cluster analysis was conducted using data for each of the aforementioned sets of input variables. In each case, participants were included if they had complete data for all included variables, because missing data for a trait has a different meaning from not having that trait. Each cluster analysis used partitioning around medoids: a more robust generalization of the k-means method.[20] The objective of using partitioning around medoids was to find a set of clusters in which the members were as similar as possible, but as dissimilar as possible to members of other clusters. The method started by

searching for k representative individuals (medoids), and the k clusters were constructed by assigning each individual to the nearest medoid. The Jaccard coefficient with 100 bootstrap samples was used to evaluate stability, as described by Hennig,[21] to assess the frequency in which pairs of participants were assigned to the same cluster across bootstraps. The clusters were numbered in an arbitrary manner. The characteristics of clusters obtained were described via summary statistics.

Prediction of cluster membership. For clusters identified with approach A, modeling was undertaken to identify predictors of cluster membership. Study patients were randomly allocated, stratifying by age, sex, and physician-assigned diagnostic label, to a training data set (70% of individuals) and a testing data set (the remaining 30% of individuals). Prediction analysis was undertaken using the maximally informative variables from input variable set A, excluding those with high loading in the cluster analysis. Gradient boosting was applied to determine the best predictors of the clusters,[22,23] and the 6 strongest variables were combined into a prediction model using the training data. Test performance was calculated for each cluster using the hold-out set to arrive at a series of binary schemes that could be subjected to receiver-operating characteristic analysis. All analyses were conducted using R 3.5.1.[24]

RESULTS
Population characteristics

Data were available for 10,885 NOVELTY patients, with almost half (46.5%) enrolled from primary care, and of whom 5163 (47.4%) had biomarker data and 7608 (70.0%) had patient-reported outcome data. Baseline characteristics are reported in detail elsewhere.[17] Overall, 52.7% of patients had a physician-assigned diagnosis of asthma alone, 12.3% asthma+COPD, and 34.9% COPD alone. Overall, 31.2%, 34.9%, and 33.7% of patients had physician-assessed mild, moderate, and severe/very severe disease, respectively.

A total of 3796 patients (mean age, 59.5 years; 54% female) had complete data for all variables selected by approach A, and 2934 patients (mean age, 60.7 years; 53% female) had complete data for all variables selected by approach B; lack of consent for blood collection was the most common reason for exclusion. There were no clinically important differences between patients included in the cluster analyses and the whole NOVELTY population, except that there were fewer non-White patients, slightly more with type 2 (T2)-high biomarkers, and a higher proportion with 1 or more exacerbation in both cluster analyses, and there were fewer current smokers with approach B (see Table E2 in this article's Online Repository at www.jaci-inpractice.org).

Cluster distribution by physician-assigned diagnosis

Using variables from approach A, 6 clusters were identified (Figure 1, A; Table I; see also Table E3 in this article's Online Repository at www.jaci-inpractice.org). The clusters had high mathematical stability (Jaccard similarities, 0.89-0.94; see Table E4 in this article's Online Repository at www.jaci-inpractice.org), but adjacent clusters overlapped in the cluster visualization (see Figure E5 in this article's Online Repository at www.jaci-inpractice.org), reflecting overlapping clinical and biological characteristics among the patients in these clusters. Analysis based on variables from approach B also identified 6 clusters (Figure 1, B; Table II; see also Table E5 in this article's

Online Repository at www.jaci-inpractice.org) with good mathematical stability (Jaccard similarities, 0.70-0.85; see Table E6 in this article's Online Repository at www.jaci-inpractice.org); again, clusters were overlapping (see Figure E6 in this article's Online Repository at www.jaci-inpractice.org), reflecting overlap of clinical and biological characteristics among the patients.

Overall, 67% and 75% of patients with a physician-assigned diagnosis of asthma alone appeared in clusters A1 to A3 and B1 to B3, respectively, and 89% and 87% of patients with a diagnosis of COPD alone appeared in clusters A4 to A6 and B4 to B6, respectively, but each of the physician-assigned diagnoses (asthma, asthma+COPD, and COPD) appeared in all the clusters (Figure 1; Tables E3 and E5). With approach A, 42% of patients with physician-assigned diagnoses of asthma+COPD were in the "asthma-like" clusters A1 and A2, whereas with approach B, only 18% of patients with physician-assigned diagnoses of asthma+COPD were in the "asthma-like" clusters of B1 and B2 (Figure 1; Table III).

Characteristics of clusters

Approach A clusters. As presented in Table I and Table E3, patients in clusters A1 and A2 were younger, with respiratory allergies, childhood onset of respiratory symptoms, well-preserved lung function, and little clinically important breathlessness (modified Medical Research Council dyspnea grade ≥2). These clusters had the highest proportion with bronchodilator responsiveness (19% and 20%, respectively). However, there were marked differences by sex (female: cluster A1, 74%; cluster A2, 31%). Cluster A3, which was predominantly female, also featured respiratory allergies, but patients were older at onset of respiratory symptoms and had more breathlessness than patients in clusters A1 and A2 despite less airflow limitation (post-bronchodilator FEV_1/forced vital capacity < lower limit of normal); almost one-quarter had anxiety or depression. Almost one-third of the participants in cluster A3 were non-White.

In clusters A4 to A6, patients were older, mostly current/ex-smokers, with less allergy history, greater airflow limitation, and more breathlessness. However, there were again marked differences by sex, with the proportion of female patients being 70%, 45%, and 24% in clusters A4, A5, and A6, respectively. Emphysema was most common in clusters A5 and A6 (32% and 30%, respectively). Cluster A6, which was predominantly male smokers/ex-smokers, had the lowest lung function and the highest prevalence of frequent productive cough, but a similar proportion had high T2 biomarkers (blood eosinophil count ≥0.3 × 10⁹/L or F$_{ENO}$ ≥ 30 parts per billion) and bronchodilator responsiveness as in the asthma-predominant clusters A1 and A2. Across the 6 clusters, there was little variation in BMI, exacerbation rate, indoor exposure to biomass fuels, blood eosinophil count, or F$_{ENO}$.

In the prediction model, the strongest predictors of cluster membership were age, weight, childhood onset of respiratory symptoms, prebronchodilator FEV_1, duration of dust/fume exposure, and number of daily medications. The area under the curve for the receiver-operating characteristic curves ranged from 0.87 (cluster A1) down to 0.71 (cluster A4; see Figure E7 in this article's Online Repository at www.jaci-inpractice.org).

Approach B clusters. As presented in Table II and Table E5, clusters B1 to B3 predominantly comprised female patients and clusters B4 to B6 predominantly comprised male
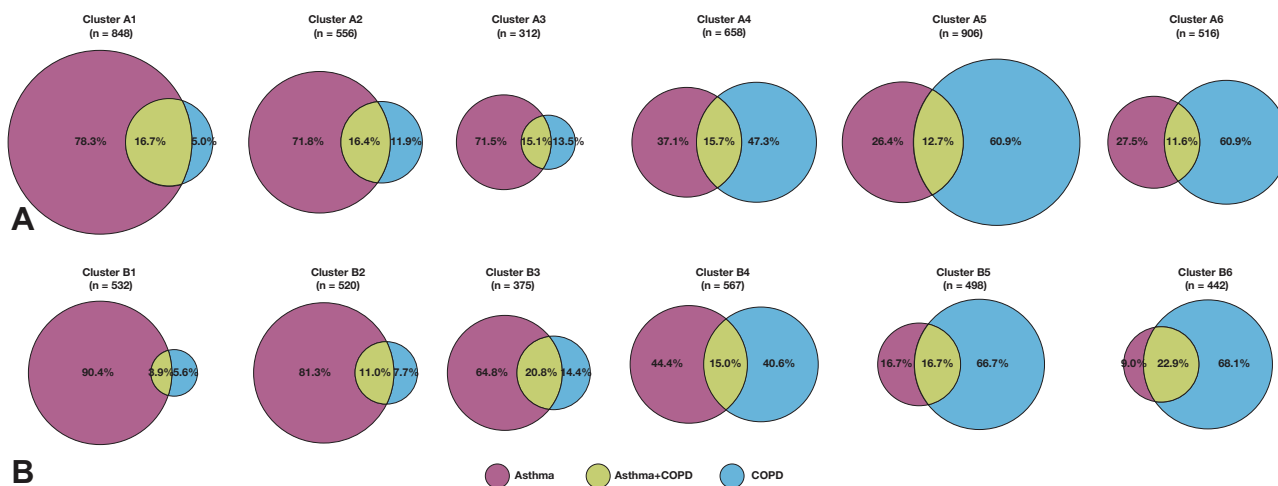
FIGURE 1. Physician-assigned diagnosis overlap by patient clusters: (**A**) approach A and (**B**) approach B to variable selection.

patients. There was some variation by ethnicity, with North-East Asian participants mostly found in clusters B1 and B4. Three clusters (B3, B5, and B6) comprised patients with high levels of breathlessness (2 of which, clusters B3 and B6, also had higher exacerbation rates) and 3 clusters (B1, B2, and B4) comprised patients with fewer symptoms and exacerbations.

The clusters were further characterized as follows: cluster B1, respiratory allergies or high-T2 markers common, with normal lung function; cluster B2, high respiratory allergies or T2 markers, more airflow limitation, with frequent productive cough in a third of patients; cluster B3, breathlessness and frequent productive cough common, airflow limitation more severe, and both blood neutrophils and T2 markers high; cluster B4, less severe airflow limitation, but high proportion of current/ex-smokers; cluster B5, more severe airflow limitation, high proportion of current/ex-smokers, and symptoms of breathlessness common; cluster B6, more severe airflow limitation, highest proportion of current/ex-smokers, very high proportions with breathlessness and frequent productive cough, and high blood neutrophils.

DISCUSSION

In the large, global, real-life NOVELTY study of patients with physician-assigned diagnoses of asthma and/or COPD recruited from primary and specialist care, cluster analyses using 2 different approaches to variable selection each found 6 identifiable but overlapping clusters, reflecting overlap of clinical and biological characteristics among patients in the clusters. Each cluster also included all 3 diagnostic labels. With both approaches, about 60% of patients with physician-diagnosed asthma appeared in 2 clusters characterized by respiratory allergies and younger age, and about 90% of patients with physician-diagnosed COPD appeared in 3 clusters characterized by current/ex-smoking and lower lung function. However, differences were seen between clusters, particularly with the nonfacilitated approach A, in features such as sex, ethnicity, breathlessness, frequent productive cough, and blood cell counts, suggesting the existence of phenotypes that differ from conventional diagnostic characteristics for asthma and COPD.

Previous studies

Several previous studies have applied cluster analysis to patients with either asthma or COPD, with substantial variation in their populations, biosampling, and variable selection. For clustering among cohorts with mild, moderate, and/or severe asthma, the Severe Asthma Research Program used clinical variables,[26] with later inclusion of FENO, blood, and bronchoscopic variables for a subpopulation,[27] and Haldar et al[28] used clinical and sputum variables. Some studies selected variables on the basis of clinical feasibility, including a Swedish population study that used questionnaire data alone[29] and the Airways Disease Endotyping for Personalized Therapeutics (ADEPT) study (which selected symptom control, airway hyperresponsiveness, and blood eosinophils),[30] or on the basis of a specific hypothesis; for example, a severe asthma registry study investigated clustering of only blood eosinophils, FENO, and IgE.[31] In patients with COPD, the ECLIPSE study,[32] the COPDGene study,[33] and other studies[34] have explored clustering approaches using clinical and physiological variables, with some studies including comorbidities, inflammatory profiles, and/or imaging. All these studies identified clusters within asthma and COPD, but because they were limited to asthma *or* COPD, they had limited ability to identify patterns that may be shared across these diagnostic labels. In addition, their populations and variables were siloed, with the asthma studies lacking data on emphysema or chronic bronchitis and the COPD studies excluding young adults and lacking data on allergies. Very few studies have used clustering approaches in patients with asthma *and/or* COPD; they include a small study in patients with severe asthma or moderate to severe COPD,[35] and a study in patients aged 40 years or older from random population samples in New Zealand and China.[36] NOVELTY is unique in this regard because it includes a large and broad spectrum of patients with a physician-assigned asthma, COPD, or asthma+COPD diagnosis recruited globally in a real-life setting from both primary and specialist care, with the same diagnosis-agnostic variables collected for all patients.

Methodologic aspects

In this analysis, we used 2 methods to select variables from among more than 170 baseline clinical features and readily

TABLE I. Distribution of key characteristics across approach A patient clusters (for full results, see Table E3)

Characteristic	Cluster A1 (n = 848)	Cluster A2 (n = 556)	Cluster A3 (n = 312)	Cluster A4 (n = 658)	Cluster A5 (n = 906)	Cluster A6 (n = 516)
Demographic						
Age, mean (y)	50.5	50.8	56.0	66.1	68.4	62.0
Age at symptom onset, mean (y)	11.4	15.6	31.5	55.8	60.3	50.0
Sex, female (%)	73.9	30.9	83.3	69.8	44.5	24.0
Ethnicity (%)						
White*	83.6	88.1	68.6	78.3	84.5	83.7
Black*	3.1	2.0	8.3	6.5	2.8	1.2
North-East Asian	5.4	4.3	11.2	8.4	7.7	5.4
South-East Asian	1.4	0.4	1.3	0.3	0.6	0.8
BMI, median (kg/m^2)	27.2	27.5	28.9	26.8	27.3	27.3
Exposure (%)						
Current smoker	12.5	15.5	12.8	17.6	20.0	24.6
Never smoked	55.4	47.1	53.5	25.8	17.5	16.9
Indoor use of biomass fuels/gas (ever)	76.8	73.7	77.9	76.0	71.6	76.2
Comorbidity (%)						
Respiratory allergies†	63.4	57.0	56.4	29.3	26.9	27.3
Emphysema	7.7	11.7	9.9	25.7	32.0	30.2
Anxiety/depression	17.6	12.2	24.0	19.8	15.3	12.0
Physiological measurements						
Post-BD FEV$_1$, % predicted, mean	81.8	79.1	82.2	71.8	70.3	67.9
Post-BD FEV$_1$/FVC < LLN (%)	34.7	39.2	29.4	45.9	51.3	57.1
Biomarker‡						
F$_{ENO}$ (ppb)	27.7	30.9	26.8	22.9	24.3	26.4
T2-high§ (%)	41.7	44.8	39.1	33.7	32.1	39.7
Blood EOS count (×10^9/L), mean ± SD	0.22 ± 0.17	0.22 ± 0.16	0.21 ± 0.19	0.21 ± 0.19	0.20 ± 0.13	0.22 ± 0.17
Blood EOS count (≥0.30) ×10^9/L (%)	23.3	21.6	18.9	18.5	16.1	20.3
Blood neutrophil count (×10^9/L), mean ± SD	4.51 ± 1.77	4.59 ± 1.73	4.61 ± 1.99	4.74 ± 1.86	4.76 ± 1.67	4.81 ± 1.67
PROs, HRU, and exacerbations						
mMRC dyspnea grade ≥2 (%)	26.9	26.4	38.8	42.2	40.0	42.1
Frequent productive cough‖ (%)	27.6	31.2	29.6	31.1	33.1	39.0
Physician-reported exacerbations (any severity) during past year, mean	0.89	1.00	0.84	0.73	0.69	0.81

BD, Bronchodilator; *EOS*, eosinophil; *FVC*, forced vital capacity; *HRU*, health care resource utilization; *LLN*, lower limit of normal; *mMRC*, modified Medical Research Council; *ppb*, parts per billion; *PRO*, patient-reported outcome; *SGRQ*, St George's Respiratory Questionnaire.
*The terms "Caucasian" and "African American" were used in the electronic case report form for recording patient ethnicity.
†Allergic rhinitis/sinusitis (seasonal or perennial) or animal/mold allergy.
‡Biosamples for analysis of biomarkers were not collected from patients in Brazil (n = 202).
§Blood EOS count ≥0.30 × 10^9/L or F$_{ENO}$ ≥30 ppb.
‖Identified from SGRQ by positive responses to both questions regarding cough and phlegm production on most or several days a week over the past 3 mo.[25]

available biomarkers in the NOVELTY data set: approach A was a hypothesis-free approach starting with reduction of variables to a low-correlation set, whereas approach B was a facilitated approach with variable selection based on adaptive grouping of clinically related traits. As might be expected, the latter showed a clearer divide between physician-assigned labels of asthma and COPD across the clusters (Figure 1; Table III), but the fact that 2 different approaches identified similar and complementary clusters strengthens the analysis. With the clinically directed approach B, patients with asthma+COPD clustered more with COPD-only diagnoses, whereas with the nonfacilitated approach A, they clustered more with asthma-only diagnoses (Table III). This may have clinical importance, because multiple studies have shown that patients with diagnoses of both asthma and COPD are more likely to die or be hospitalized if they are treated as if for COPD with long-acting bronchodilators alone (without inhaled

corticosteroids), whereas this is not found in patients with COPD alone.[37,38]

Interpretation of results

Overall, each approach resulted in 3 clusters with features including younger age and respiratory allergies (with many patients having an asthma diagnosis) and 3 clusters with features including older age at onset of respiratory symptoms, current/ex-smoking, clinically important breathlessness, more emphysema diagnoses, and often greater airflow limitation (with many patients having a diagnosis of COPD). However, the differences seen between clusters and between approaches in the distribution of several variables, including sex, age, age at onset, ethnicity, breathlessness, frequent productive cough, and T2 biomarkers suggest phenotypes that may differ from conventional diagnostic characteristics. The marked differences by sex between clusters in

TABLE II. Distribution of selected variables across approach B patient clusters (for full results, see Table E5)

Characteristic	Cluster B1 (n = 532)	Cluster B2 (n = 520)	Cluster B3 (n = 375)	Cluster B4 (n = 567)	Cluster B5 (n = 498)	Cluster B6 (n = 442)
Demographic						
Age, mean (y)	51.5	51.4	61.0	66.2	68.2	66.9
Age at symptom onset, mean (y)	27.4	26.0	31.5	49.8	49.9	48.2
Sex, female (%)	69.9	65.6	70.4	38.3	35.9	40.7
Ethnicity (%)						
White*	73.1	82.9	78.9	72.0	82.7	89.6
Black*	1.5	2.9	3.5	2.1	1.6	2.0
North-East Asian	21.6	5.4	5.3	22.4	8.8	3.8
South-East Asian	0.8	1.0	1.6	0.2	0.2	0.2
BMI, median (kg/m^2)	24.4	28.9	30.5	26.4	26.8	28.5
Exposure (%)						
Current smoker	5.6	11.7	10.1	16.2	17.5	21.3
Never smoked	68.2	56.2	58.4	18.7	9.0	2.0
Indoor use of biomass fuels/gas (ever)	70.1	79.0	74.4	76.5	72.1	72.2
Comorbidity (%)						
Respiratory allergies†	65.6	73.3	47.2	28.9	22.1	21.7
Emphysema	3.0	5.4	11.5	23.6	41.0	46.6
Anxiety/depression	10.5	19.0	20.3	10.6	13.1	18.8
Physiological measurements						
Post-BD FEV$_1$, % predicted, mean	99.1	88.5	63.9	87.9	56.3	51.6
Post-BD FEV$_1$/FVC < LLN (%)	15.7	23.6	53.8	32.5	73.5	77.0
Biomarker‡						
F$_{ENO}$ (ppb)	29.4	33.2	31.7	28.0	21.0	18.9
T2-high§ (%)	42.9	49.8	53.1	39.2	28.5	25.1
Blood EOS count (×10^9/L), mean ± SD	0.19 ± 0.15	0.24 ± 0.20	0.26 ± 0.22	0.20 ± 0.14	0.20 ± 0.18	0.19 ± 0.14
Blood EOS count (≥0.30 ×10^9/L) (%)	17.7	26.0	30.9	15.3	15.7	15.6
Blood neutrophil count (×10^9/L), mean ± SD	3.82 ± 1.46	4.54 ± 1.57	5.47 ± 2.11	4.13 ± 1.39	4.98 ± 1.74	5.09 ± 1.89
PROs, HRU, and exacerbations						
mMRC dyspnea grade ≥2 (%)	7.3	23.7	61.7	10.7	40.8	71.6
Frequent productive cough‖ (%)	14.5	36.4	53.3	12.5	22.8	55.4
Physician-reported exacerbations (any severity) during past year, mean	0.41	0.88	1.83	0.25	0.69	1.24

BD, Bronchodilator; *EOS*, eosinophil; *FVC*, forced vital capacity; *HRU*, health care resource utilization; *LLN*, lower limit of normal; *mMRC*, modified Medical Research Council; *ppb*, parts per billion; *PRO*, patient-reported outcome; *SGRQ*, St George's Respiratory Questionnaire.
*The terms "Caucasian" and "African American" were used in the electronic case report form for recording patient ethnicity.
†Allergic rhinitis/sinusitis (seasonal or perennial) or animal/mold allergy.
‡Biosamples for analysis of biomarkers were not collected from patients in Brazil (n = 202).
§Blood EOS count ≥0.30 × 10^9/L or F$_{ENO}$ ≥30 ppb.
‖Identified from SGRQ by positive responses to both questions regarding cough and phlegm production on most or several days a week over the past 3 mo.[25]

approach A are of particular interest because sex was not included following hypothesis-free testing to select variables that captured most variation in the data. However, the variability caused by differences between sexes was captured by other selected variables: prebronchodilator FEV$_1$, smoking status, and weight. Possible drivers of differences between sexes in obstructive lung disease range from risk factors, genetics, pathophysiology, and presentation of disease to treatment and response.[39] These cross-sectional cluster analyses were the first step toward the primary objective of NOVELTY, which is to identify clinical phenotypes and molecular endotypes on the basis of biomarkers and/or clinical parameters that are associated with differential outcomes for symptom burden, clinical evolution, and health care utilization over time. The fact that these initial cross-sectional clusters based on conventional clinical features and readily available biomarkers were overlapping (eg, as visualized in Figures E5 and

E6) indicates overlap of clinical and biological characteristics between members in adjacent clusters. Clearly, therefore, these clusters do not reflect discrete underlying mechanisms, but ongoing analysis toward this aim will include identification of individual patient trajectories in these variables over 3 to 5 years of follow-up, and deep "omics" analysis of the stored biobank.

The present findings suggest opportunities for precision medicine: for example, there are many potential underlying mechanisms for frequent productive cough (one of the most common traits found in NOVELTY[25]), and high blood eosinophils or F$_{ENO}$ may be either rapidly responsive to inhaled corticosteroids[40] or, in severe asthma, may be persistently elevated despite systemic corticosteroids.[41] The ultimate goal is precision medicine, based on identifying specific molecular mechanisms; this has been extremely successful in oncology but remains challenging for complex chronic diseases such as obstructive lung

TABLE III. Proportions of each diagnosis and category of severity, by cluster (row percentages)

Approach A	Cluster A1 (n = 848)	Cluster A2 (n = 556)	Cluster A3 (n = 312)	Cluster A4 (n = 658)	Cluster A5 (n = 906)	Cluster A6 (n = 516)
Diagnosis group at baseline, n (%)						
Asthma	664 (34.7)	399 (20.9)	223 (11.7)	244 (12.8)	239 (12.5)	142 (7.4)
Asthma+COPD	142 (25.4)	91 (16.3)	47 (8.4)	103 (18.5)	115 (20.6)	60 (10.8)
COPD	42 (3.2)	66 (5.0)	42 (3.2)	311 (23.4)	552 (41.6)	314 (23.7)
Physician-assessed severity (%)						
Mild	23.9	13.6	9.7	17.1	23.1	12.7
Moderate	21.5	14.5	8.0	20.4	24.6	11.1
Severe/very severe	21.7	15.8	7.1	14.7	23.9	16.8
Approach B	**Cluster B1 (n = 532)**	**Cluster B2 (n = 520)**	**Cluster B3 (n = 375)**	**Cluster B4 (n = 567)**	**Cluster B5 (n = 498)**	**Cluster B6 (n = 442)**
Diagnosis group at baseline, n (%)						
Asthma	481 (31.6)	423 (27.8)	243 (16.0)	252 (16.6)	83 (5.5)	40 (2.6)
Asthma+COPD	21 (4.9)	57 (13.4)	78 (18.4)	85 (20.0)	83 (19.5)	101 (23.9)
COPD	30 (3.0)	40 (4.1)	54 (5.5)	230 (23.3)	332 (33.6)	301 (30.5)
Physician-assessed severity (%)						
Mild	30.2	19.3	5.3	30.0	11.5	3.7
Moderate	17.9	19.7	10.9	21.7	16.4	13.3
Severe/very severe	7.0	14.3	21.7	6.9	22.8	27.4

disease. In the meantime, the present findings emphasize the heterogeneity of asthma, asthma+COPD, and COPD, and the importance of avoiding a siloed, restrictive approach to research.

For clinical practice, there is increasing interest in identifying specific treatable traits or modifiable risk factors and comorbidities that may direct specific treatment choices, independent of the diagnostic label, with a recent analysis describing the distribution and patterns of 30 treatable traits in the large NOVELTY population.[42]

Strengths and potential limitations

Strengths of this analysis include NOVELTY's wide geographic coverage and large population of patients with asthma and/or COPD from both primary and specialist care. The inclusion criteria were very broad, compared with the narrow eligibility criteria for regulatory clinical trials.[8-11] Clinical, physiological, and biomarker variables were measured in a standardized manner, and the value of this approach was emphasized by the finding of respiratory allergies, emphysema, and frequent productive cough across the conventional diagnostic labels. Potential limitations are that there were few young patients in NOVELTY (18% were <45 years) and that patients with missing data for any of the included variables were excluded from the analyses. Reasons for missing data include lack of patient/regulator consent for blood sampling for biomarker analysis and failure of some patients to complete patient-reported outcome questionnaires; the resulting reduction in the number of patients included in the analyses may affect the representativeness of the findings. In addition, most patients in NOVELTY were White (74.7%), which limits the ability to generalize the findings to other ethnicities. However, country and ethnicity were included in the initial set of 79 variables considered for input into the cluster analysis for approach A, but neither was ranked among the strongest drivers of variation in the data. In addition (as seen in Table E2), there were no major systematic differences between patients included versus the whole NOVELTY population, except that there were fewer non-White patients, slightly more

with T2-high biomarkers, and a higher proportion with 1 or more exacerbation in both cluster analyses, and there were fewer current smokers with approach B.

CONCLUSIONS

These analyses in the NOVELTY population, based on facilitated and nonfacilitated selection of variables, each identified 6 overlapping clusters that crossed the conventional diagnostic labels of asthma, COPD, or asthma+COPD and revealed several discriminatory features that differed from conventional diagnostic characteristics. The overlap between clusters, reflecting the overlapping clinical and biological characteristics of patients included in these clusters, suggests that they do not reflect discrete underlying mechanisms. These findings emphasize the importance of ongoing research to identify specific underlying molecular endotypes and potential treatment targets across the spectrum of obstructive lung disease.

article's Online Repository at www.jaci-inpractice.org. Medical writing support, under the direction of the authors, was provided by Richard Knight (PhD), CMC Connect, a division of IPG Health Medical Communications, funded by AstraZeneca in accordance with Good Publication Practice 2022 guidelines.[43]

REFERENCES

1. GBD 2017 Disease and Injury Incidence and Prevalence Collaborators. Global, regional, and national incidence, prevalence, and years lived with disability for 354 diseases and injuries for 195 countries and territories, 1990-2017: a systematic analysis for the Global Burden of Disease Study 2017. Lancet 2018;392:1789-858.
2. Fletcher CM. Terminology, definitions, and classification of chronic pulmonary emphysema and related conditions: a report of the conclusions of a Ciba guest symposium. Thorax 1959;14:286-99.
3. Pavord ID, Beasley R, Agusti A, Anderson GP, Bel E, Brusselle G, et al. After asthma: redefining airways diseases. Lancet 2018;391:350-400.
4. Vanfleteren LEGW, Kocks JWH, Stone IS, Breyer-Kohansal R, Greulich T, Lacedonia D, et al. Moving from the Oslerian paradigm to the post-genomic era: are asthma and COPD outdated terms? Thorax 2014;69:72-9.
5. Bateman ED, Reddel HK, van Zyl-Smit RN, Agustí A. The asthma-COPD overlap syndrome: towards a revised taxonomy of chronic airways diseases? Lancet Respir Med 2015;3:719-28.
6. Uchida A, Sakaue K, Inoue H. Epidemiology of asthma-chronic obstructive pulmonary disease overlap (ACO). Allergol Int 2018;67:165-71.
7. Milne S, Mannino D, Sin DD. Asthma-COPD overlap and chronic airflow obstruction: definitions, management, and unanswered questions. J Allergy Clin Immunol Pract 2020;8:483-95.
8. Travers J, Marsh S, Caldwell B, Williams M, Aldington S, Weatherall M, et al. External validity of randomized controlled trials in COPD. Respir Med 2007; 101:1313-20.
9. Travers J, Marsh S, Williams M, Weatherall M, Caldwell B, Shirtcliffe P, et al. External validity of randomised controlled trials in asthma: to whom do the results of the trials apply? Thorax 2007;62:219-23.
10. Brown T, Jones T, Gove K, Barber C, Elliott S, Chauhan A, et al. Randomised controlled trials in severe asthma: selection by phenotype or stereotype. Eur Respir J 2018;52:1801444.
11. Pahus L, Burgel PR, Roche N, Paillasseur JL, Chanez P. Randomized controlled trials of pharmacological treatments to prevent COPD exacerbations: applicability to real-life patients. BMC Pulm Med 2019;19:127.
12. Agusti A, Bel E, Thomas M, Vogelmeier C, Brusselle G, Holgate S, et al. Treatable traits: toward precision medicine of chronic airway diseases. Eur Respir J 2016;47:410-9.
13. Poon LKM, Zhang NL, Chen T, Wang Y. Variable selection in model-based clustering: to do or to facilitate. Presented at: 27th International Conference on Machine Learning, Haifa, Israel, 21−24 June 2010;47:887-94. Accessed June 6, 2023. https://icml.cc/Conferences/2010/papers/26.pdf
14. Grira N, Crucianu M, Boujemaa N. Unsupervised and semi-supervised clustering: a brief survey. Accessed June 6, 2023. http://cedric.cnam.fr/~crucianm/src/BriefSurveyClustering.pdf
15. Deliu M, Sperrin M, Belgrave D, Custovic A. Identification of asthma subtypes using clustering methodologies. Pulm Ther 2016;2:19-41.
16. Reddel HK, Gerhardsson de Verdier M, Agustí A, Anderson G, Beasley R, Bel EH, et al. Prospective observational study in patients with obstructive lung disease: NOVELTY design. ERJ Open Res 2019;5:00036-2018.
17. Reddel HK, Vestbo J, Agustí A, Anderson GP, Bansal AT, Beasley R, et al. Heterogeneity within and between physician-diagnosed asthma and/or COPD: NOVELTY cohort. Eur Respir J 2021;58:2003927.
18. Cadima J, Cerdeira JO, Minhoto M. Computational aspects of algorithms for variable selection in the context of principal components. Comput Stat Data Anal 2004;47:225-36.
19. Harrell FE. Multivariable modeling strategies. In: Harrell FE, editor. Regression modeling strategies: with applications to linear models, logistic regression, and survival analysis. New York: Springer; 2001. p. 53-85.
20. Kaufman L, Rousseeuw PJ. Clustering by means of medoids. In: Reports of the Faculty of Mathematics and Informatics. Delft, The Netherlands: Faculty of Mathematics and Informatics, Delft University of Technology; 1987.
21. Hennig C. Cluster-wise assessment of cluster stability. Comput Stat Data Anal 2007;52:258-71.
22. Friedman JH. Greedy function approximation: a gradient boosting machine. Ann Stat 2001;29:1189-232.
23. Friedman JH. Stochastic gradient boosting. Comput Stat Data Anal 2002;38: 367-78.
24. R Core Team. R: a language and environment for statistical computing. 2018. Accessed January 13, 2021. https://www.r-project.org/
25. Hughes R, Rapsomaniki E, Janson C, Keen C, Make BJ, Burgel PR, et al. Frequent productive cough: symptom burden and future exacerbation risk among patients with asthma and/or COPD in the NOVELTY study. Respir Med 2022;200:106921.
26. Moore WC, Meyers DA, Wenzel SE, Teague WG, Li H, Li X, et al. Identification of asthma phenotypes using cluster analysis in the Severe Asthma Research Program. Am J Respir Crit Care Med 2010;181:315-23.
27. Wu W, Bleecker E, Moore W, Busse WW, Castro M, Chung KF, et al. Unsupervised phenotyping of Severe Asthma Research Program participants using expanded lung data. J Allergy Clin Immunol 2014;133:1280-8.
28. Haldar P, Pavord ID, Shaw DE, Berry MA, Thomas M, Brightling CE, et al. Cluster analysis and clinical asthma phenotypes. Am J Respir Crit Care Med 2008;178:218-24.
29. Kisiel MA, Zhou X, Sundh J, Ställberg B, Lisspers K, Malinovschi A, et al. Data-driven questionnaire-based cluster analysis of asthma in Swedish adults. NPJ Prim Care Respir Med 2020;30:14.
30. Loza MJ, Djukanovic R, Chung KF, Horowitz D, Ma K, Branigan P, et al. Validated and longitudinally stable asthma phenotypes based on cluster analysis of the ADEPT study. Respir Res 2016;17:165.
31. Denton E, Price DB, Tran TN, Canonica GW, Menzies-Gow A, FitzGerald JM, et al. Cluster analysis of inflammatory biomarker expression in the international severe asthma registry. J Allergy Clin Immunol Pract 2021;9:2680-8.e7.
32. Rennard SI, Locantore N, Delafont B, Tal-Singer R, Silverman EK, Vestbo J, et al. Identification of five chronic obstructive pulmonary disease subgroups with different prognoses in the ECLIPSE cohort using cluster analysis. Ann Am Thorac Soc 2015;12:303-12.
33. Castaldi PJ, Dy J, Ross J, Chang Y, Washko GR, Curran-Everett D, et al. Cluster analysis in the COPDGene study identifies subtypes of smokers with distinct patterns of airway disease and emphysema. Thorax 2014;69:415-22.
34. Nikolaou V, Massaro S, Fakhimi M, Stergioulas L, Price D. COPD phenotypes and machine learning cluster analysis: a systematic review and future research agenda. Respir Med 2020;171:106093.
35. Ghebre MA, Pang PH, Diver S, Desai D, Bafadhel M, Haldar K, et al. Biological exacerbation clusters demonstrate asthma and chronic obstructive pulmonary disease overlap with distinct mediator and microbiome profiles. J Allergy Clin Immunol 2018;141:2027-2036.e12.
36. Fingleton J, Huang K, Weatherall M, Guo Y, Ivanov S, Bruijnzeel P, et al. Phenotypes of symptomatic airways disease in China and New Zealand. Eur Respir J 2017;50:1700957.
37. Suissa S, Ernst P. Observational studies of inhaled corticosteroid effectiveness in COPD: lessons learned. Chest 2018;154:257-65.
38. Kendzerska T, Aaron SD, To T, Licskai C, Stanbrook M, Vozoris NT, et al. Effectiveness and safety of inhaled corticosteroids in older individuals with chronic obstructive pulmonary disease and/or asthma. A population study. Ann Am Thorac Soc 2019;16:1252-62.
39. Somayaji R, Chalmers JD. Just breathe: a review of sex and gender in chronic lung disease. Eur Respir Rev 2022;31:210111.
40. Heaney LG, Busby J, Bradding P, Chaudhuri R, Mansur AH, Niven R, et al. Remotely monitored therapy and nitric oxide suppression identifies nonadherence in severe asthma. Am J Respir Crit Care Med 2019;199: 454-64.
41. Graff S, Vanwynsberghe S, Brusselle G, Hanon S, Sohy C, Dupont LJ, et al. Chronic oral corticosteroids use and persistent eosinophilia in severe asthmatics from the Belgian severe asthma registry. Respir Res 2020;21:214.
42. Agustí A, Rapsomaniki E, Beasley R, Hughes R, Müllerová H, Papi A, et al. Treatable traits in the NOVELTY study. Respirology 2022;27:929-40.
43. DeTora LM, Toroser D, Sykes A, Vanderlinden C, Plunkett FJ, Lane T, et al. Good Publication Practice (GPP) guidelines for company-sponsored biomedical research: 2022 update. Ann Intern Med 2022;175:1298-304.